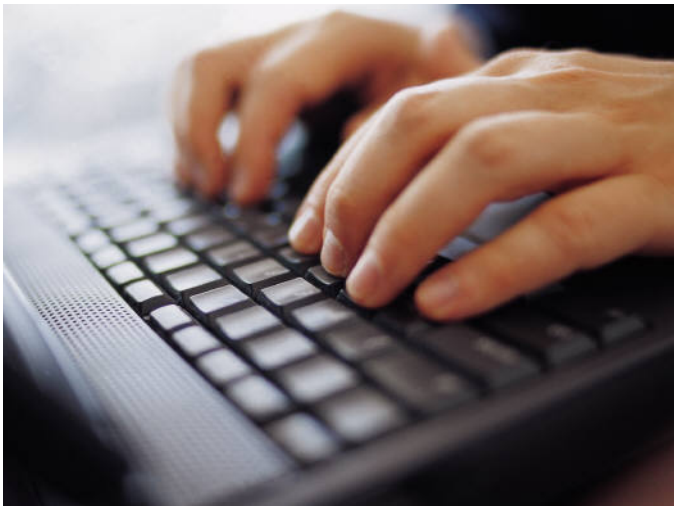
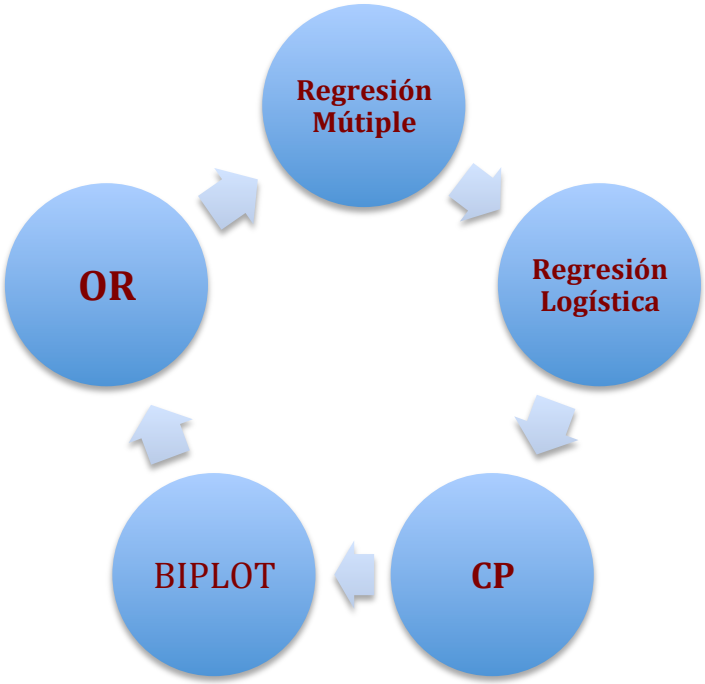


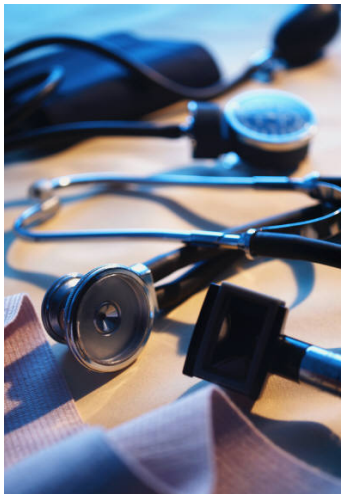
BIOESTADÍSTICA AVANZADA

2013



Mª Purificación Galindo Villardón
pgalindo@usal.es

Dpto. Estadísica
Universidad de Salamanca



ESTADÍSTICA DESCRIPTIVA

Mª PURIFICACIÓN GALINDO VILLARDÓN
pgalindo@usal.es

Tema 0: Revisión de conceptos

Medidas de tendencia central — Medidas de dispersión

MEDIA $\bar{x} = \frac{\sum_i x_i}{n}$

Recorrido $R = Máx - Mín$

Varianza $s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$

Desviación Típica $s = +\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$

Coeficiente de variación $C.V. = \frac{s}{|\bar{x}|}$

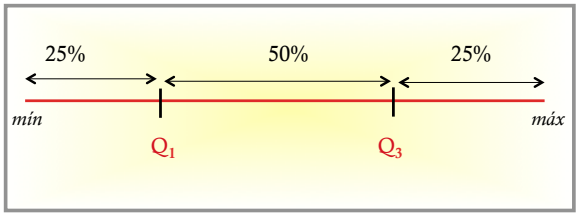
Error Estándar

$ESM = S / \sqrt{n}$

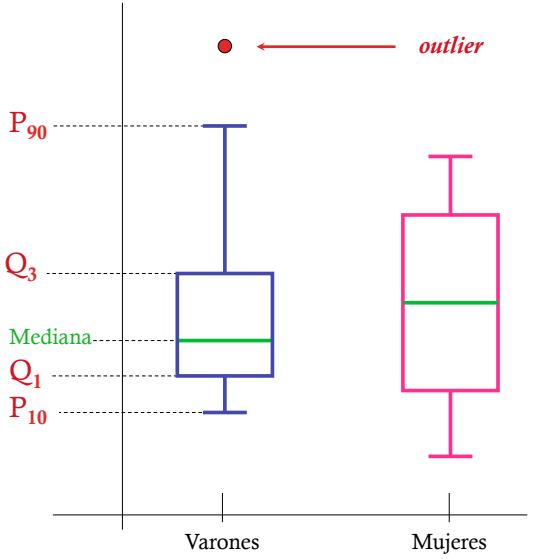
Medidas de tendencia central — Medidas de dispersión

MEDIANA — Recorrido Intercuartílico

$R.I. = Q_3 - Q_1$



Medidas de tendencia central — Gráfico Box-Plot



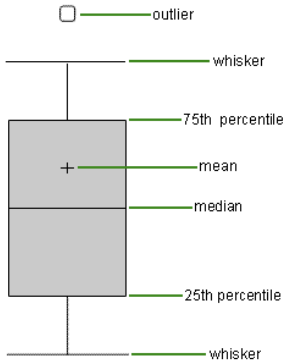
Citation Reports

actor Box Plot

pact Factor box plot depicts the distribution of Impact Factors for all journals in the category. The horizontal line that forms the tc he 75th percentile. The horizontal line that forms the bottom is the 25th percentile. The horizontal line that intersects the box is t Factor for the category. The cross represents the mean value.

ıtal lines above and below the box represent maximum and minimum values that are no more than 1.5 times the span of the inte which is the range of values between the 25th and the 75th percentiles. These lines are commonly referred to as "whiskers."

n circle represents an outlier, which is a single value greater or less than the extremes indicated by the whiskers.



INFERENCIA

¿Cuánto vale la media de una población?

$\mu = ?$

A partir de los datos

$\bar{x} = 30$ → *Estimación puntual* → $\mu = 30$

Estimación por intervalo

ESM=0.5

$\bar{x} - 1ESM \le \mu \le \bar{x} + 1ESM$ → $29.5 \le \mu \le 30.5$ **Con una confianza del 68%**

$\bar{x} - 2ESM \le \mu \le \bar{x} + 2ESM$ → $28.5 \le \mu \le 31.5$ **Con una confianza del 99%**

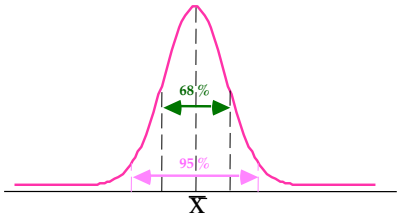
INTERVALO CONFIANZA

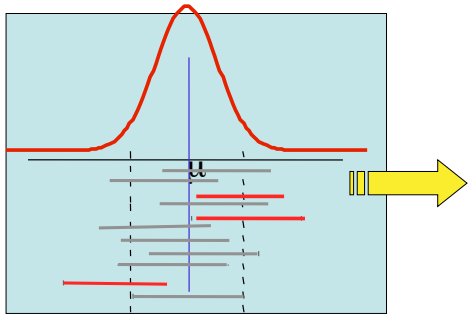
$$I_{\mu}^{1-\alpha} = \bar{x} \pm Z_{\alpha/2} \frac{\hat{s}}{\sqrt{n}}$$

dispersión

tamaño muestra

Nivel de confianza





El 95% de los intervalos contiene la verdadera media de la población.
El 5% no lo contienen

Mas información en

TABLAS DE CONTINGENCIA

M. P. Galindo Villardón
pgalindo@usal.es
DEPARTAMENTO ESTADÍSTICA
Universidad de Salamanca

- **HIPOTESIS DE PARTIDA: H_0**
  Las dos variables en estudio son independientes

- **HIPOTESIS ALTERNATIVA: Ha**
  Las dos variables en estudio están relacionadas

Un ejemplo:
Tabla de frecuencias observadas

	Peor	*Igual*	*Mejor*	TOTAL
Trat1	7	28	115	150
Trat2	15	20	85	120
Trat3	10	30	90	130
Trat4	5	40	115	160
TOTAL	37	118	405	560

¿Cómo se contrasta?

- Partimos de una tabla de frecuencias observadas
- Se calculan las frecuencias que cabría esperar si las dos variables fueran independientes

Ho ⇕

fe_{ij} = (Total fila i-ésima) (Total columna j-ésima) / Total global

Nuestro ejemplo:
Tabla de frecuencias observadas

	Peor	*Igual*	*Mejor*	TOTAL
Trat1	**7**			**150**
Trat2				
Trat3				
Trat4				
TOTAL	**37**			**560**

$fo_{11} = 7$ $fe_{11} = (150x37)/560 = 9.91$

TABLA DE FRECUENCIAS ESPERADAS

	Peor	**Igual**	**Mejor**	**TOTAL**
Trat 1	9,91	31,61	108,48	150
Trat 2	7,93	25,28	86,79	120
Trat 3	8,59	27,39	94,02	130
Trat 4	10,57	33,72	115,71	160
TOTAL	37	118	405	560

CÓMO MEDIR LAS DISCREPANCIAS

•Se calcula la diferencia entre ambas magnitudes (**fo_{ij} - fe_{ij}**), para todas y cada una de las casillas de la tabla.

Estadígrafo de contraste

$$\chi^2_{exp} = \sum_i \sum_j \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

fo_{ij} = frecuencia observada para la ij-ésima casilla.

fe_{ij} = frecuencia esperada para la ij-ésima casilla.

· **Rechazaremos H_0 cuando** χ^2 experimental > χ^2 crítico

GRADOS DE LIBERTAD: (Nº filas-1) (Nº Columnas -1)

En el ordenador, al lado del valor experimental, (suma de discrepancias entre frecuencias observadas y esperadas), aparece el p-valor

El p-valor nos indica el riesgo que corremos al rechazar la Ho (independencia) después de haber explorado los datos.

Si el p-valor es menor de 0.05, rechazamos H_0 y aceptamos la Ha. Si p-valor > 0.05, NO

CÁLCULO DEL VALOR EXPERIMENTAL

$$\chi^2_{exp} = \sum_i \sum_j \frac{\left(fo_{ij} - fe_{ij}\right)^2}{fe_{ij}}$$

➢ *Aplicado a nuestro ejemplo el resultado sería:*

$$\chi^2_{exp} = \frac{(7-9.91)^2}{9.91} + \ldots + \frac{(115-115.71)^2}{115.71} = 13.87$$

¿Qué concluimos?

TABLA DE LA JI-CUADRADO

	0.9950	0.9750	0.950	0.900	0.200	0.10	0.050	0.025	0.010	0.001
1	0.0000393	0.000982	0.00393	0.0158	1.642	2.706	3.841	5.024	6.635	10.828
2	0.010	0.0506	0.103	0.211	3.219	4.605	5.991	7.378	9.510	13.816
3	0.0717	0.216	0.352	0.584	4.642	6.251	7.851	9.348	11.345	16.266
4	0.207	0.484	0.711	1.064	5.989	7.779	9.488	1.143	13.277	18.467
5	0.412	0.831	1.145	1.610	7.289	9.236	11.070	12.833	15.086	20.515
6	0.676	1.237	1.635	2.204	8.558	10.645	12.59	14.449	16.812	22.458
7	0.989	1.690	2.167	2.833	9.803	12.017	14.067	16.013	18.475	24.322
8	1.344	2.180	2.733	3.490	11.030	13.362	15.507	17.535	20.090	26.124
9	1.735	2.700	3.325	4.168	12.242	14.684	16.919	19.023	21.666	27.877
10	2.156	3.247	3.940	4.865	13.442	15.987	18.307	20.483	23.209	29.588
11	2.603	3.816	4.575	5.578	14.631	17.275	19.675	21.920	24.725	31.264
12	3.074	4.404	5.226	6.304	15.812	18.549	21.026	23.337	26.217	32.909
13	3.565	5.009	5.892	7.042	16.985	19.812	22.362	24.736	27.688	34.528
14	4.075	5.629	6.571	7.790	18.151	21.064	23.685	26.119	29.141	36.123
15	4.601	6.262	7.261	8.547	19.311	22.307	24.996	27.488	30.578	37.697
16	5.142	6.908	7.962	9.312	20.465	23.452	26.296	28.845	32.000	39.252
17	5.697	7.564	8.672	10.085	21.615	24.769	27.587	30.191	33.409	40.790
18	6.265	8.231	9.390	10.865	22.760	25.989	28.869	31.526	34.805	42.312
19	6.844	8.907	10.117	11.651	23.900	27.204	30.144	32.852	36.191	43.820
20	7.434	9.591	10.851	12.443	25.038	28.412	31.410	34.170	37.566	45.315
21	8.034	10.283	11.591	13.240	26.171	29.615	32.671	35.479	38.932	46.979
22	8.643	10.982	12.338	14.041	27.301	30.813	33.924	36.781	40.289	48.268
23	9.260	11.689	13.091	14.848	28.429	32.007	35.172	38.076	41.638	49.728
24	9.886	12.401								51.179
25	10.520									52.620
26	11.160									54.052
27	11.808		16.151	18.114	32.912	36.741	40.113	43.195	46.963	55.476
28	12.461	15.308	16.928	18.939		37.916	41.337	44.461	48.278	56.892
29	13.121	16.047	17.708	19.769	35.139	39.087	42.557	45.722	49.588	58.301
30	13.787	16.791	18.493	20.599	36.250	40.256	43.773	46.979	50.892	59.703

$\chi^2_{0.05,6} = 12.59$

χ^2_{exp} **13.87 > 12.59 => p-valor <0.05**

Tablas poco ocupadas:

Las tablas de contingencia con frecuencias bajas o nulas llevan a cometer error de tipo 1.

Este tipo de tablas produce una frecuencia teórica muy baja, que distorsiona el resultado y lleva a rechazar Ho.

Frecuencias relativas /porcentajes:

Las tablas de contingencia trabajadas sobre porcentajes pueden llevar asociadas aceptaciones indebidas de la Hipótesis nula; es decir pueden llevarnos a suponer que dos variables son independientes cuando en realidad están relacionadas.
(Incremento en el Riesgo Tipo II)

Problemática de trabajar con frecuencias relativas /porcentajes:

Las tablas de contingencia trabajadas sobre porcentajes pueden llevar asociadas aceptaciones indebidas de la Hipótesis nula; es decir pueden llevarnos a suponer que dos variables son independientes cuando en realidad están relacionadas.
(Incremento en el Riesgo Tipo II)

Y ahora con el SPSS...

¿Cómo meter los datos?

Datos

- **Como tabla de contingencia ya construida.**

	Peor	*Igual*	*Mejor*	
Trat1	7	28	115	
Trat2	15	20	85	
Trat3	10	30	90	
Trat4	5	40	115	
				560

TABLA.sav - SPSS Editor de datos

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana

15 : DIAGNÓSTICO

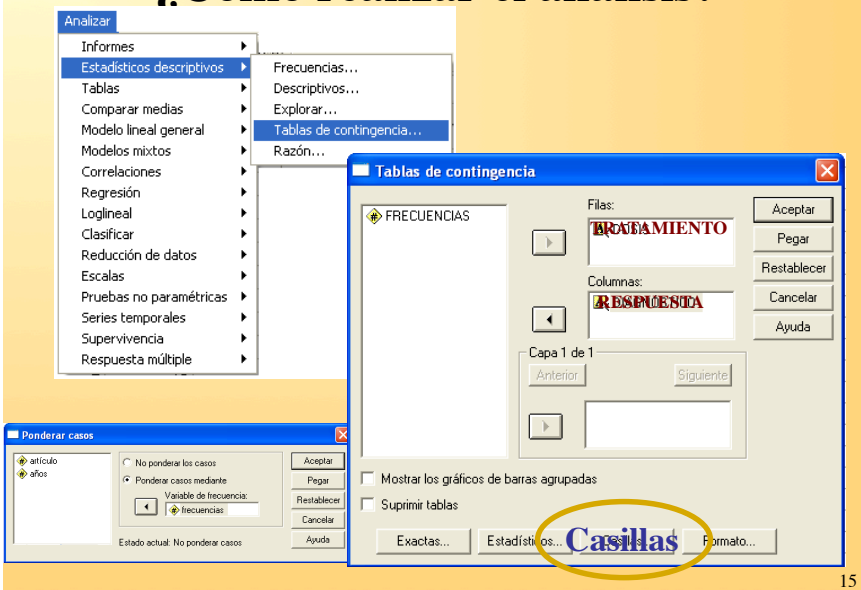
	DOSIS	DIAGNÓSTICO	FRECUENCIAS
1	TRATAM_1	1_PEOR	7
2	TRATAM_1	2_IGUAL	28
3	TRATAM_1	3_MEJOR	115
4	TRATAM_2	1_PEOR	15
5	TRATAM_2	2_IGUAL	20
6	TRATAM_2	3_MEJOR	85
7	TRATAM_3	1_PEOR	10
8	TRATAM_3	2_IGUAL	30
9	TRATAM_3	3_MEJOR	90
10	TRATAM_4	1_PEOR	5
11	TRATAM_4	2_IGUAL	40

Tabla de contingencia DOSIS * DIAGNÓSTICO

Recuento

		DIAGNÓSTICO			
		1_PEOR	2_IGUAL	3_MEJOR	Total
DOSIS	TRATAM_1	7	28	115	150
	TRATAM_2	15	20	85	120
	TRATAM_3	10	30	90	130
	TRATAM_4	5	40	115	160
Total		37	118	405	560

¿Cómo realizar el análisis?



Casillas

Tabla de contingencia DOSIS * DIAGNÓSTICO

Recuento

		DIAGNÓSTICO			
		1_PEOR	2_IGUAL	3_MEJOR	Total
DOSIS	TRATAM_1	7	28	115	150
	TRATAM_2	15	20	85	120
	TRATAM_3	10	30	90	130
	TRATAM_4	5	40	115	160
Total		37	118	405	560

Frecuencias o Recuentos

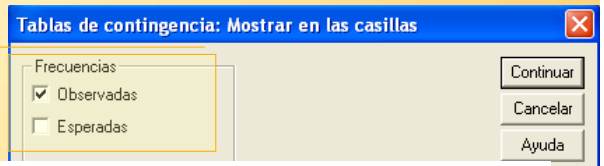


Tabla de contingencia DOSIS * DIAGNÓSTICO

Frecuencia esperada

		DIAGNÓSTICO			
		1_PEOR	2_IGUAL	3_MEJOR	Total
DOSIS	TRATAM_1	9,9	31,6	108,5	150,0
	TRATAM_2	7,9	25,3	86,8	120,0
	TRATAM_3	8,6	27,4	94,0	130,0
	TRATAM_4	10,6	33,7	115,7	160,0
Total		37,0	118,0	405,0	560,0

Casillas

Porcentajes

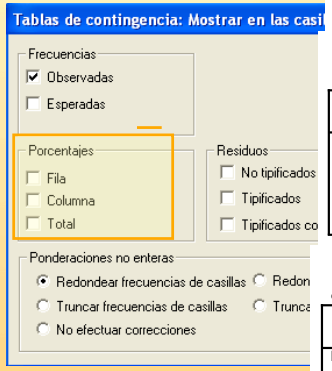


Tabla de contingencia DOSIS * DIAGNÓSTICO

% de DOSIS

		DIAGNÓSTICO			
		1 PEOR	2 IGUAL	3 MEJOR	Total
DOSIS	TRATAM_1	4,7%	18,7%	76,7%	100,0%
	TRATAM_2	12,5%	16,7%	70,8%	100,0%
	TRATAM_3	7,7%	23,1%	69,2%	100,0%
	TRATAM_4	3,1%	25,0%	71,9%	100,0%
Total		6,6%	21,1%	72,3%	100,0%

Tabla de contingencia DOSIS * DIAGNÓSTICO

% de DIAGNÓSTICO

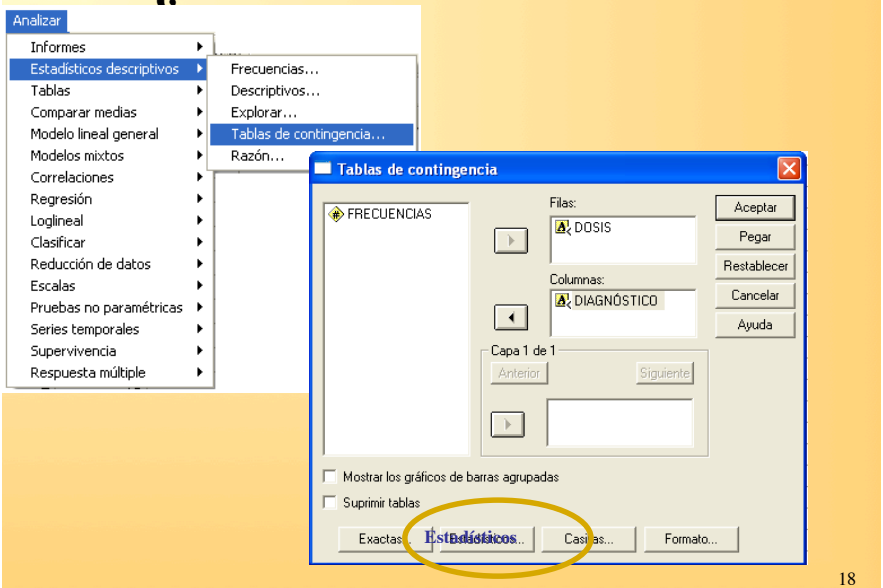
		DIAGNÓSTICO			
		1 PEOR	2 IGUAL	3 MEJOR	Total
DOSIS	TRATAM_1	18,9%	23,7%	28,4%	26,8%
	TRATAM_2	40,5%	16,9%	21,0%	21,4%
	TRATAM_3	27,0%	25,4%	22,2%	23,2%
	TRATAM_4	13,5%	33,9%	28,4%	28,6%
Total		100,0%	100,0%	100,0%	100,0%

Tabla de contingencia DOSIS * DIAGNÓSTICO

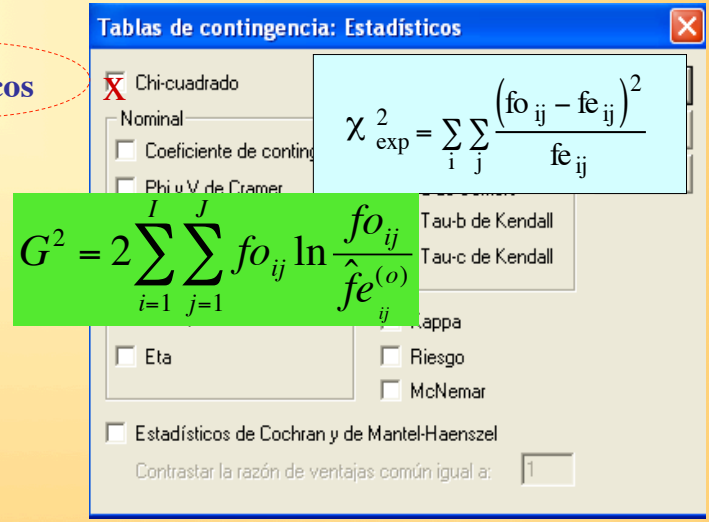
% del total

		DIAGNÓSTICO			
		1 PEOR	2 IGUAL	3 MEJOR	Total
DOSIS	TRATAM_1	1,3%	5,0%	20,5%	26,8%
	TRATAM_2	2,7%	3,6%	15,2%	21,4%
	TRATAM_3	1,8%	5,4%	16,1%	23,2%
	TRATAM_4	,9%	7,1%	20,5%	28,6%
Total		6,6%	21,1%	72,3%	100,0%

¿Cómo realizar el análisis?



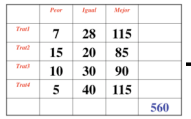
Estadísticos



$$\chi^2_{exp} = \sum_i \sum_j \frac{\left(fo_{ij} - fe_{ij}\right)^2}{fe_{ij}}$$

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} fo_{ij} \ln \frac{fo_{ij}}{\hat{fe}_{ij}^{(o)}}$$

Ambos siguen una distribución Chi cuadrado con (I-1) (J-1) gardos de libertad



	Peor	Igual	Mejor	
Total	7	28	115	
Total	15	20	85	
Total	10	30	90	
Total	5	40	115	
				560

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	13,871[a]	6	,031
Razón de verosimilitud	13,378	6	,037
N de casos válidos	560		

P-valor

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 7,93.

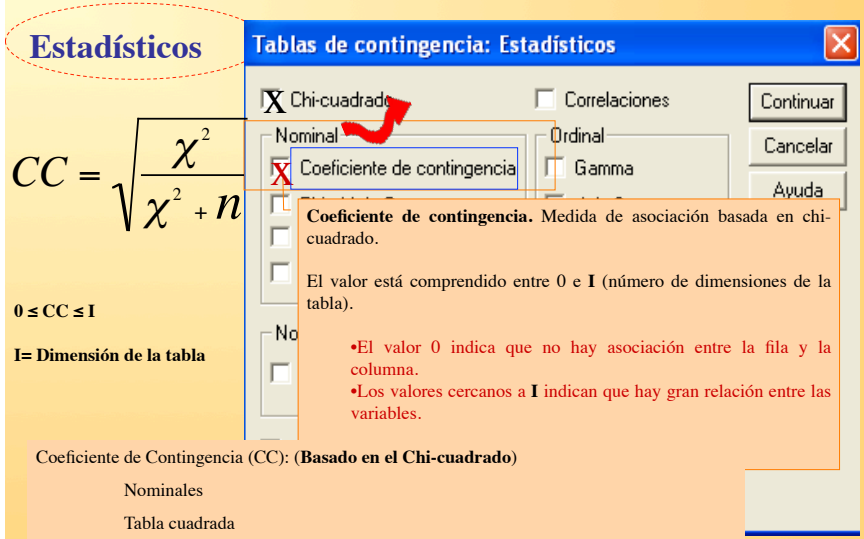
En el ordenador, al lado del valor experimental, aparece el p-valor

El p-valor nos indica el riesgo que corremos al rechazar la Ho (independencia) después de haber explorado los datos.

Si el p-valor es menor de 0.05, rechazamos H**0** y aceptamos la Ha. Si p-valor > 0.05, NO

GRADO DE ASOCIACIÓN

Estadísticos



$$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

0 ≤ CC ≤ I

I= Dimensión de la tabla

Coeficiente de contingencia. Medida de asociación basada en chi-cuadrado.

El valor está comprendido entre 0 e **I** (número de dimensiones de la tabla).

- El valor 0 indica que no hay asociación entre la fila y la columna.
- Los valores cercanos a **I** indican que hay gran relación entre las variables.

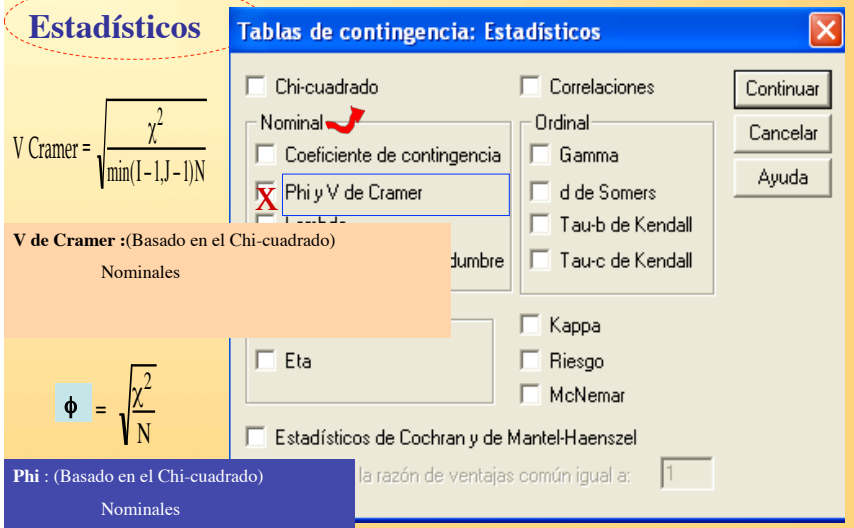
Coeficiente de Contingencia (CC): (**Basado en el Chi-cuadrado**)

Nominales

Tabla cuadrada

Rol simétrico

Estadísticos



$$V\ Cramer = \sqrt{\frac{\chi^2}{\min(I-1,J-1)N}}$$

V de Cramer :(Basado en el Chi-cuadrado)

Nominales

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Phi : (Basado en el Chi-cuadrado)

Nominales

Dicotómicas

Casillas

Residuos

Tabla de contingencia DOSIS * DIAGNÓSTICO

Residuo

		DIAGNÓSTICO		
		1 PEOR	2 IGUAL	3 MEJOR
DOSIS	TRATAM_1	-2,9	-3,6	6,5
	TRATAM_2	7,1	-5,3	-1,8
	TRATAM_3	1,4	2,6	-4,0
	TRATAM_4	-5,6	6,3	-,7

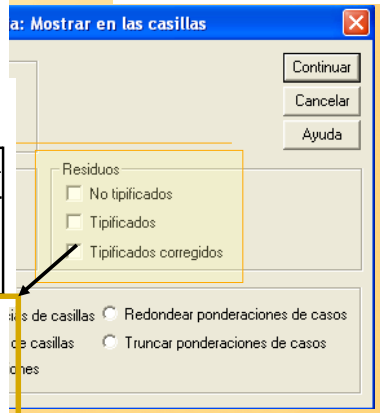
Residuos tipificados

		DIAGNÓSTICO		
		1 PEOR	2 IGUAL	3 MEJOR
DOSIS	TRATAM_1	-,9	-,6	,6
	TRATAM_2	2,5	-1,1	-,2
	TRATAM_3	,5	,5	-,4
	TRATAM_4	-1,7	1,1	-,1

Residuos corregidos

		DIAGNÓSTICO		
		1 PEOR	2 IGUAL	3 MEJOR
DOSIS	TRATAM_1	-1,1	-,8	1,4
	TRATAM_2	2,9	-1,3	-,4
	TRATAM_3	,6	,6	-,9
	TRATAM_4	-2,1	1,4	-,1

	Peor	Igual	Mejor	
Trat1	7	28	115	
Trat2	15	20	85	
Trat3	10	30	90	
Trat4	5	40	115	
				560



Mas información en

Universidad de Salamanca
Departamento de ESTADÍSTICA

M. P. Galindo Villardón
pgalindo@usal.es

Contrastes de hipótesis

1. ***Hipótesis nula*** (H_0), ***Hipótesis alternativa*** (H_1)
2. ***Nivel de significación*** (α)
3. ***Estadístico de contraste*** (Z, t, ...)
4. ***Región crítica y región de aceptación*** (RC y RA)
5. ***Conclusiones*** (estadísticas y no estadísticas: médicas, biológicas, económicas, etc.)

Contrastes de hipótesis

➢ ***Nivel de significación*** (α)

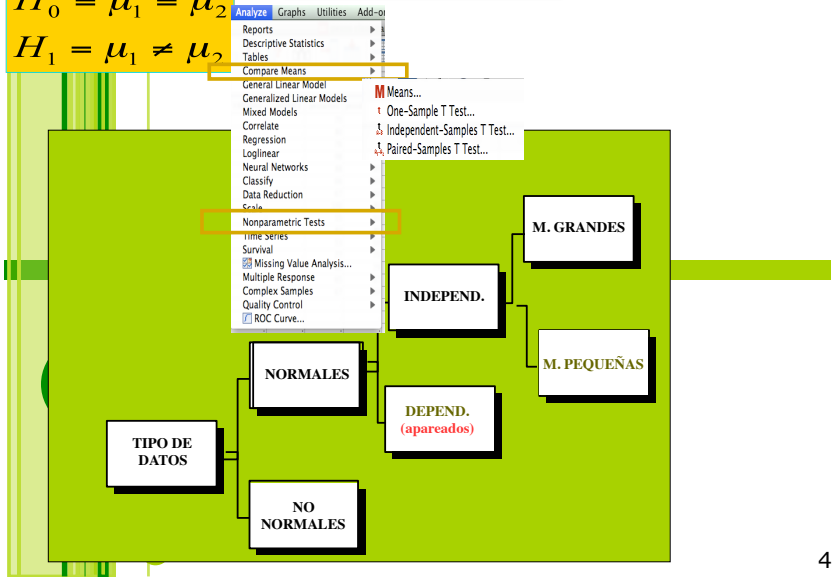
Probabilidad de cometer *error tipo I*, es decir, probabilidad de rechazar la hipótesis nula siendo cierta. Habitualmente 1%, 5%.

Resultado de la prueba estadística	Hipótesis cierta: H_0 Inocente	Hipótesis cierta: H_1 Culpable
Rechazo Ho (A la cárcel...)	Error tipo I (α)	Correcto (Potencia (1-β))
Acepto Ho (En libertad...)	Correcto	Error tipo II (β)

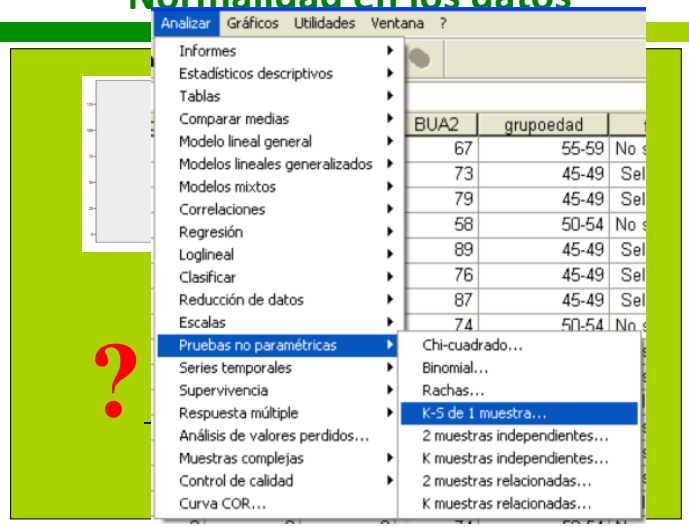
CONTRASTE PARA LA IGUALDAD DE MEDIDAS DE TENDENCIA CENTRAL

$H_0 = \mu_1 = \mu_2$

$H_1 = \mu_1 \neq \mu_2$



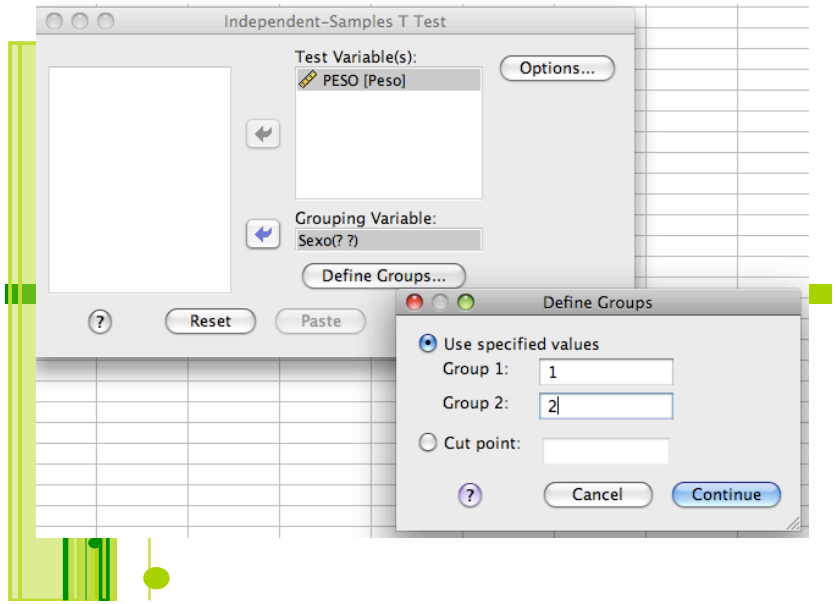
Normalidad en los datos



Sexo-peso

DATOS

	INDEPENDIENTES					APAREADOS	
	Sexo	Peso	VAR00001	VAR00002	VAR00003	pesoantes	pesodespues
1	varón	70,00	.	.		70	65
2	varón	68,00	.	.		80	82
3	varón	72,00	.	.		76	70
4	varón	71,00	.	.		98	88
5	varón	70,00	.	.		88	89
6	mujer	56,00	.	.		66	665
7	varón	65,00	.	.		67	60
8	mujer	60,00	.	.		100	80
9	mujer	45,00	.	.		55	60
10	mujer	55,00	.	.		82	80
11	varón	56,00	.	.		53	50
12	mujer	66,00	.	.		77	75
13	mujer	55,00	.	.		75	79
14	mujer	44,00	.	.		65	67
15	varón	66,00	.	.		67	60
16	mujer	56,00	.	.		.	.
17	mujer	44,00	.	.		.	.
18	varón	70,00	.	.		.	.
19	varón	80,00	.	.		.	.
20	varón	75,00	.	.		.	.
21	varón	67,00	.	.		.	.
22	mujer	55,00	.	.		.	.
23	varón	66,00	.	.		.	.
24	mujer	50,00	.	.		.	.




```
T-TEST GROUPS=Sexo(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=Peso
  /CRITERIA=CI(.9500).
```

T-Test

[DataSet0]

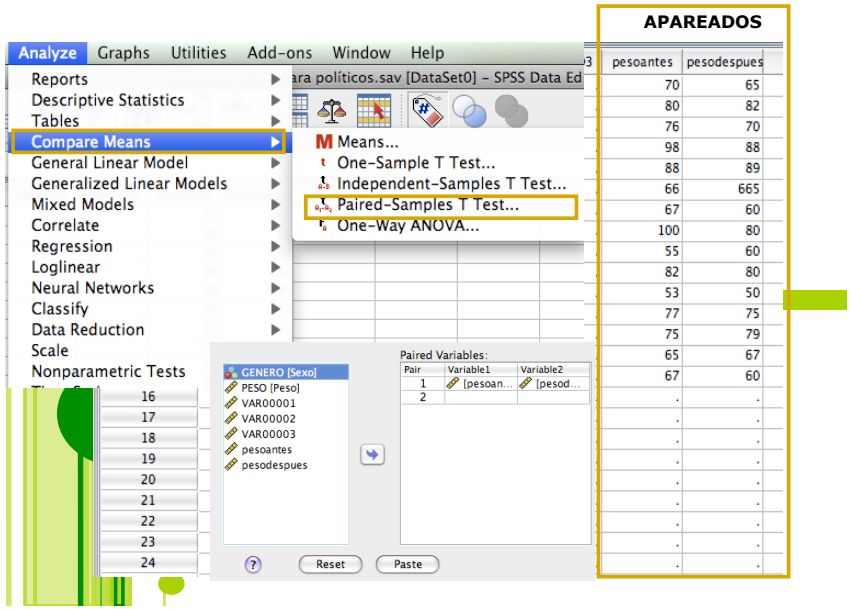
Group Statistics

	GENERO	N	Mean	Std. Deviation	Std. Error Mean
PESO	varón	13	68,9231	5,63414	1,56263
	mujer	11	53,2727	6,94393	2,09367

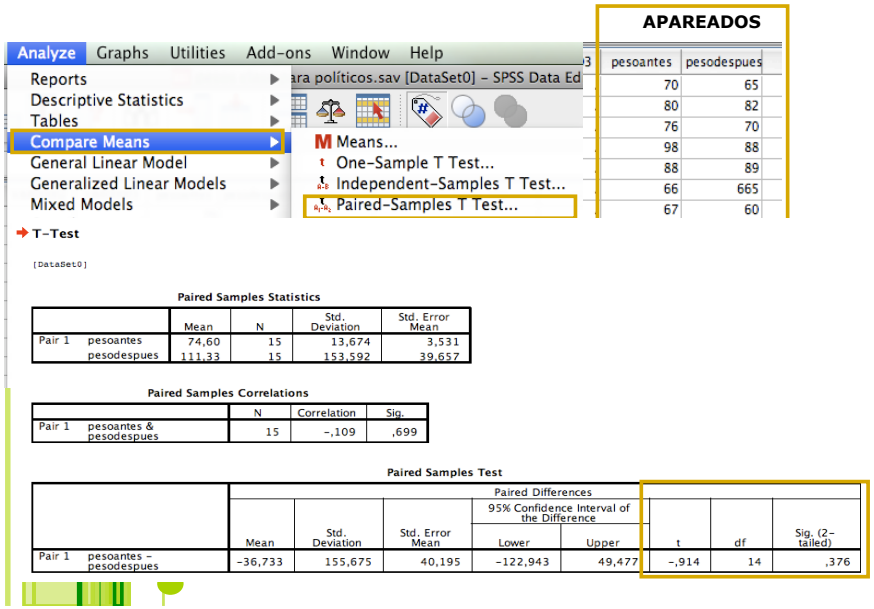
Independent Samp

		Levene's Test for Equality of Variances: F	Sig.	Ho: $\mu_1=\mu_2$ t-test for Equality of Means: t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Ho: $\sigma_1=\sigma_2$								
PESO	Equal variances assumed	,936	,344	6,099	22	,000	15,65035	2,56601
	Equal variances not assumed			5,991	19,263	,000	15,65035	2,61252

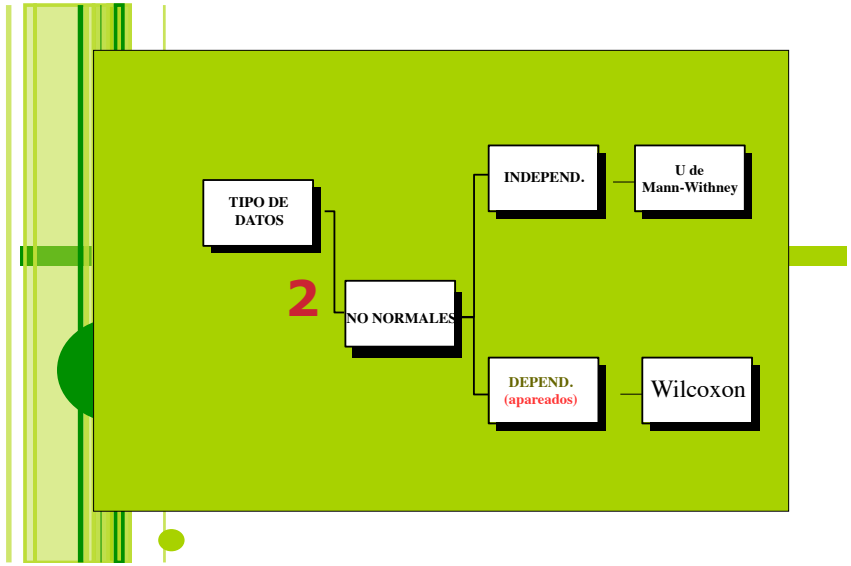
DATOS



DATOS



CONTRASTE PARA LA IGUALDAD DE MEDIDAS DE TENDENCIA CENTRAL: MEDIANAS



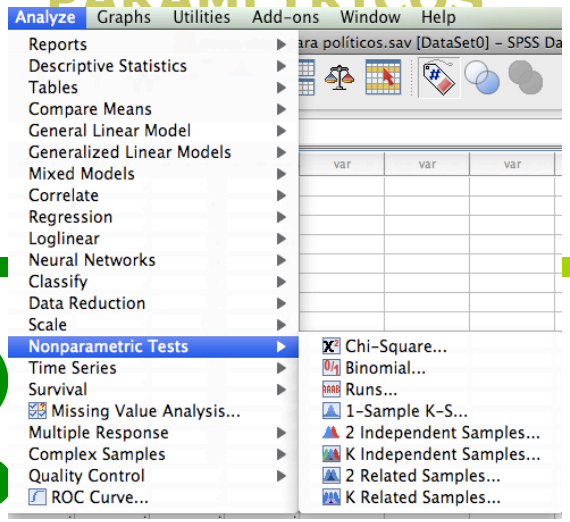
Tests **NO PARAMETRICOS**

Datos NO NORMALES

- **U de Mann- Withney**
- **Wilcoxon**

- **Comparan MEDIANAS**
- **Trabajan sobre rangos de ord**
- **Son menos potentes**

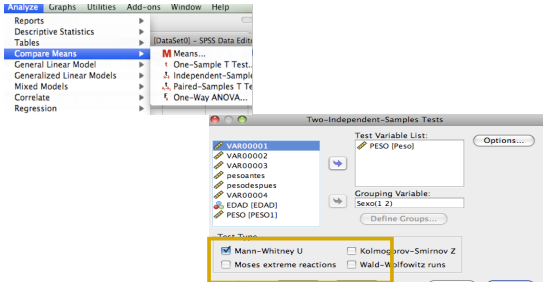
CONTRASTES NO PARAMÉTRICOS



Sexo-peso

INDEPENDIENTES

	Sexo	Peso	VARO
1	varón	70,00	
2	varón	68,00	
3	varón	72,00	
4	varón	71,00	
5	varón	70,00	
6	mujer	56,00	
7	varón	65,00	
8	mujer	60,00	
9	mujer	45,00	
10	mujer	55,00	
11	varón	56,00	
12	mujer	66,00	
13	mujer	55,00	
14	mujer	44,00	
15	varón	66,00	
16	mujer	56,00	
17	mujer	44,00	
18	varón	70,00	
19	varón	80,00	
20	varón	75,00	
21	varón	67,00	
22	mujer	55,00	
23	varón	66,00	
24	mujer	50,00	



NPar Tests

[DataSet0] /Users/puri/Desktop/pesos clase para políticos.sav

Mann-Whitney Test

Ranks

	GE...	N	Mean Rank	Sum of Ranks
PESO	varón	13	17,62	229,00
	mujer	11	6,45	71,00
	Total	24		

Test Statistics

	PESO
Mann-Whitney U	5,000
Wilcoxon W	71,000
Z	-3,867
Asymp. Sig. (2-tailed)	,000
Exact Sig. [2*(1-tailed...	,000

a. Not corrected for ties.

b. Grouping Variable: GENERO

ANOVA

Análisis de la varianza

Dra. Purificación Galindo Villardón
pgalindo@usal.es

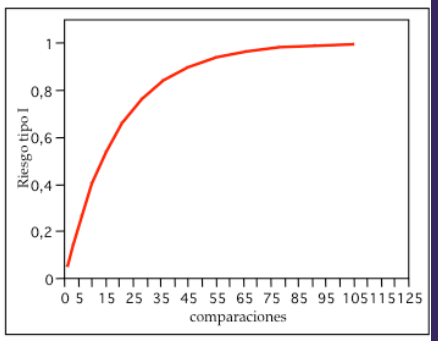
Y si hay más de dos Grupos…?

- ***Fijemos α = probabilidad de rechazo indebido de Ho***

$$\alpha = P(H_0 \equiv \mu_1 = \mu_2 = \mu_3 \text{ cierta, pero rechazamos algún de los contrastes por parejas}) =$$
$$= 1 - P(H_0 \equiv \mu_1 = \mu_2 = \mu_3 \text{ cierta y aceptamos todos los contrastes por parejas}) = 1 - 0.95^3 = \mathbf{0.1426} \neq 0.05$$

T de STUDENT

grupos	comparaciones	riesgo tipo I
2	1	0,05
3	3	0,1426
4	6	0,2649
5	10	0,4013
6	15	0,5367
7	21	0,6594
8	28	0,7622
9	36	0,8422
10	45	0,9006
11	55	0,9405
12	66	0,9661
13	78	0,9817
14	91	0,9906
15	105	0,9954



ANOVA

ANOVA

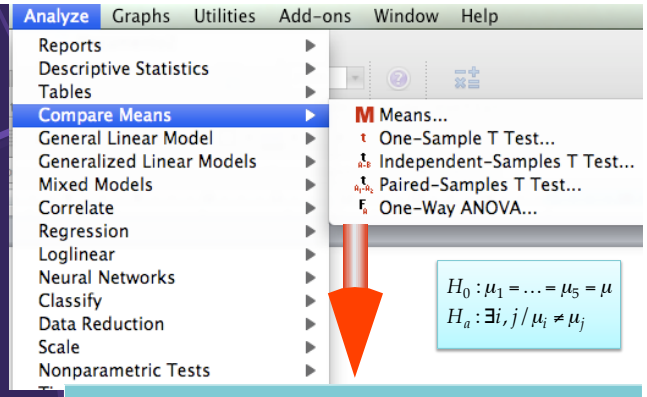
Tratamiento				
D1	D2	D3	D4	D5
1,53	3,15	2,89	3,89	3,86
1,61	3,96	2,68	3,64	3,46
3,75	3,59	4,70	5,36	3,69
2,89	1,89	4,62	3,33	4,49
3,26	1,45	4,79	6,82	5,81
	1,56	4,33	3,26	7,03
			5,10	5,49
				6,98

$$H_0 : \mu_1 = \ldots = \mu_5 = \mu$$
$$H_a : \exists i, j / \mu_i \neq \mu_j$$

Se compara la variabilidad dentro de los grupos con la variabilidad entre los grupos y se buscan los valores críticos con una F de Snedecor.

TRAT	Valores
1	1,53
1	1,61
1	3,75
1	2,89
1	3,26
2	3,15
2	3,96
2	3,59
2	1,89
2	1,45
2	1,56
3	2,89
3	2,68
3	4,70
3	4,62
3	4,79
3	4,33
4	3,89
4	3,64
4	5,36
4	3,33
4	6,82
4	3,26
4	5,10
5	3,86
5	3,46
5	3,69
5	4,49
5	5,81
5	7,03
5	5,49
5	6,98

CIA ESTADÍST



$H_0 : \mu_1 = \ldots = \mu_5 = \mu$
$H_a : \exists i, j / \mu_i \neq \mu_j$

ANOVA				
Fuente	*Suma de cuadrados*	*g.l.*	*Estimador*	*Fexp*
Entre	Q_E =32,494	4	8,123	5,540
Residual	Q_R =39,501	27	1,466	*p-valor =0,0022*
Total	Q =72,085	31		**

ANOVA

Tratamiento				
D1	D2	D3	D4	D5
1,53	3,15	2,89	3,89	3,86
1,61	3,96	2,68	3,64	3,46
3,75	3,59	4,70	5,36	3,69
2,89	1,89	4,62	3,33	4,49
3,26	1,45	4,79	6,82	5,81
	1,56	4,33	3,26	7,03
			5,10	5,49
				6,98

$H_0 : \mu_1 = \ldots = \mu_5 = \mu$
$H_a : \exists i, j / \mu_i \neq \mu_j$

ANOVA				
Fuente	*Suma de cuadrados*	*g.l.*	*Estimador*	*Fexp*
Entre	Q_E =32,494	4	8,123	5,540
Residual	Q_R =39,501	27	1,466	*p-valor =0,0022*
Total	Q =72,085	31		**

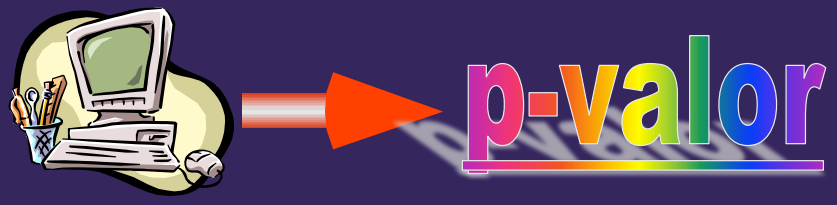
Tratamiento				
D1	D2	D3	D4	D5
1,53	3,15	2,89	3,89	3,86
1,61	3,96	2,68	3,64	3,46
3,75	3,59	4,70	5,36	3,69
2,89	1,89	4,62	3,33	4,49
3,26	1,45	4,79	6,82	5,81
	1,56	4,33	3,26	7,03
			5,10	5,49
				6,98

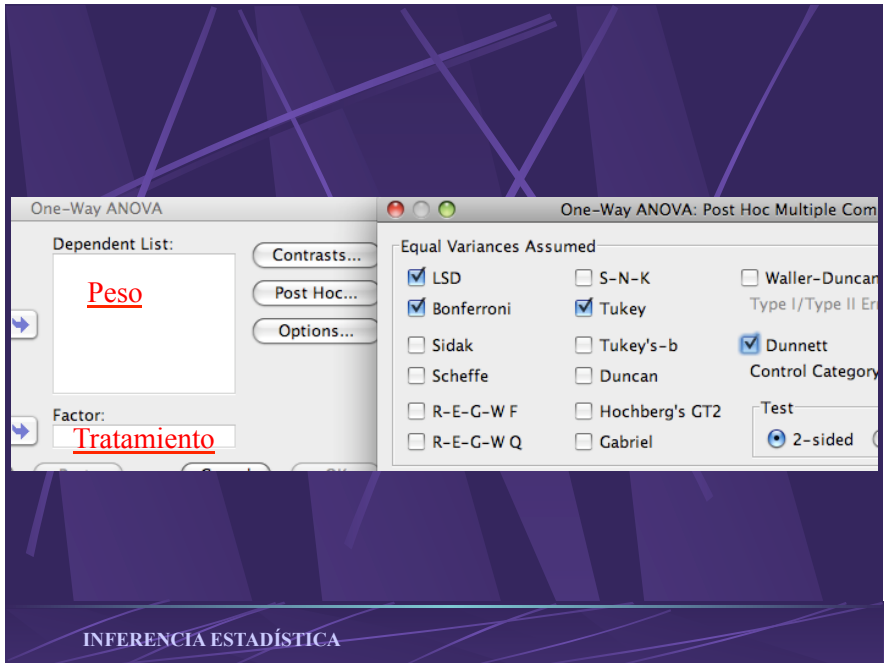
TRAT	Valores
1	1,53
1	1,61
1	3,75
1	2,89
1	3,26
2	3,15
2	3,96
2	3,59
2	1,89
2	1,45
2	1,56
3	2,89
3	2,68
3	4,70
3	4,62
3	4,79
3	4,33
4	3,89
4	3,64
4	5,36
4	3,33
4	6,82
4	3,26
4	5,10
5	3,86
5	3,46
5	3,69
5	4,49
5	5,81
5	7,03
5	5,49
5	6,98

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
TRAT	Numeric	8	0	TRATAMIENTO	{1, D1}...	None	8	Right	Ordinal
Valores	Numeric	8	2	RESPUESTA	None	None	8	Right	Scale

ANOVA

Se compara la variabilidad dentro de los grupos con la variabilidad entre los grupos y se buscan los valores críticos con una F de Snedecor.





Tests tras ANOVA

$$t = \frac{\bar{x}_{i\bullet} - \bar{x}_{i'\bullet}}{\sqrt{S_R^2\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}} \xrightarrow{Ho\ es\ cierta} t_{N-r}$$

$$Ho: \mu_i = \mu_{i'} \quad (i \neq i'; i,i' = 1,\cdots,r)$$

$$Ha: \mu_i \neq \mu_{i'}$$

$\alpha_{TUKEY} = \alpha / r$ r = nº grupos

$\alpha_{BONFERRONI} = \alpha / [r(r-1)/2\ ]$

$\alpha_{DUNNETT} = \alpha/(r-1)$

Más conservador (penaliza más)
Más difícil encontrar diferencias
Mayor protección frente al error tipo I

ANOVA DE DOS VIAS CON INTERACCIÓN

Enfermedad	*Droga* A	B	C
Esquizofrénicos	8	10	8
	4	8	6
	0	6	4
Depresivos	14	4	15
	10	2	12
	6	0	9

EJEMPLO

Fuente	Suma de cuadrados	g.l.	Estimador	F_{exp}	P-VALOR
Filas	18	**1**	18	2,04	0,1789
Columnas	48	**2**	24	2,72	0,1063
Interacción	144	**2**	72	8,15	0,0058
Residual	106	**12**	8,833		
Total	316	**17**			

El Teorema de BAYES

Análisis de una tabla tetracórica

M. P. Galindo Villardón

pgalindo@usal.es

DEPARTAMENTO ESTADÍSTICA
Universidad de Salamanca

EL TEOREMA DE BAYES EN EL CONTEXTO CLÍNICO

A ≡ Presencia de una enfermedad
$\bar{A}$ ≡ Ausencia de una enfermedad

P(A) ≡ PREVALENCIA
(Probabilidad, a priori, de la enfermedad A en la población)
P($\bar{A}$) ≡ Probabilidad de no tener la enfermedad

$\bar{B}$ ≡ Resultado positivo de un test diagnóstico.
B ≡ Resultado negativo de un test diagnóstico.

P(B/A) ≡ **SENSIBILIDAD** del test
(Probabilidad de que el test sea positivo en presencia de la enfermedad). **VP**

P($\bar{B}$/A) ≡ Probabilidad de un falso negativo **FN**

P($\bar{B}$/$\bar{A}$) ≡ **ESPECIFICIDAD** del test **VN**
(Probabilidad de obtener resultado negativo en sanos).

P(B/$\bar{A}$) ≡ Probabilidad de un falso positivo **FP**

¿P(A/B)? = VALOR PREDICTIVO POSITIVO (VPP)

¿P($\bar{A}$/$\bar{B}$) ? = VALOR PREDICTIVO NEGATIVO (VPN)

El TEOREMA DE BAYES permite calcular:

La **probabilidad de que el individuo esté enfermo en el caso de que el resultado del test diagnóstico sea positivo**:

$$P(A/B) = \frac{P(B/A).P(A)}{P(B/A).P(A) + P(B/\bar{A}).P(\bar{A})}$$

Thomas BAYES (1702 - 1761)

La **probabilidad de que el individuo esté sano condicionado a que el resultado del test diagnóstico sea negativo**:

$$P(\bar{A}/\bar{B}) = \frac{P(\bar{B}/\bar{A}).P(\bar{A})}{P(\bar{B}/\bar{A}).P(\bar{A}) + P(\bar{B}/A).P(A)}$$

Valor predictivo del test con resultado positivo:

$$= \frac{\text{prevalencia} \,.\, \text{sensibilidad}}{\text{prevalencia} \,.\, \text{sensibilidad} + (1 - \text{prev.}) \,.\, (1 - \text{espec.})}$$

Valor predictivo del test con resultado negativo:

$$= \frac{(1 - \text{prevalencia}) \,.\, \text{especificidad}}{(1 - \text{prevalencia}) \,.\, \text{especificidad} + \text{prev}\,(1 - \text{sensib.})}$$

Problema:

En una campaña de erradicación de la tuberculosis se somete la población escolar a la prueba de tuberculina. Se sabe que:

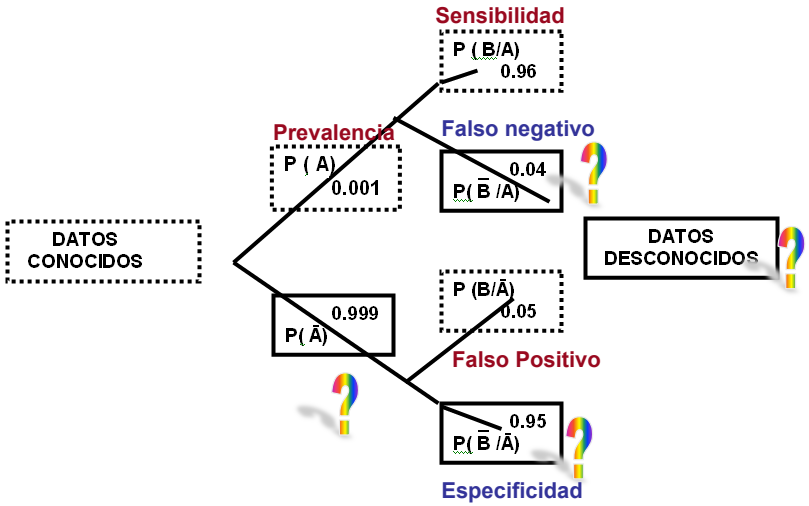
la **Sensibilidad del test es 0.96**

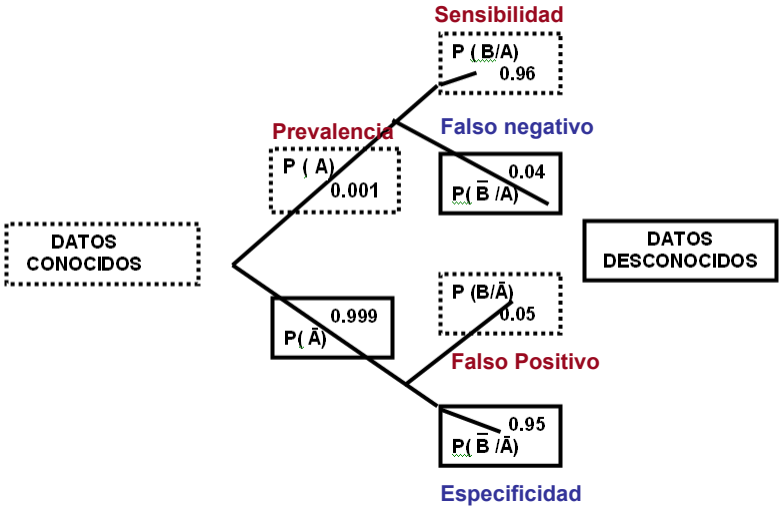
la Probabilidad de que el test sea positivo en un individuo sano es 0.05.

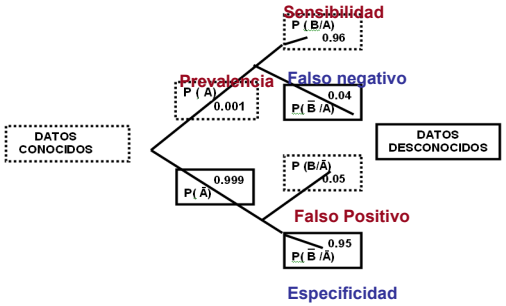
Si la **prevalencia de la enfermedad es 0.001**,

DETERMINAR:

a) **Valor predictivo para casos positivos**.
b) **Valor predictivo para casos negativos**.







$$P(A/B) = \frac{P(B/A).P(A)}{P(B/A).P(A) + P(B/\overline{A}).P(\overline{A})} =$$

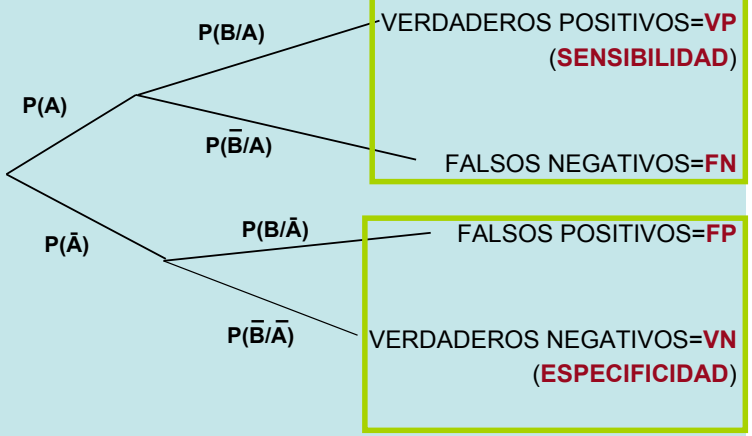
$$= \frac{0{,}96\,.\,0{,}001}{0{,}96\,.\,0{,}001 + 0{,}05\,.\,0{,}999} = VPP = 0{,}0188$$

PARA CONFIRMAR UN DIAGNÓSTICO

¿La prueba más sensible?
¿O la más específica?

¿Y para DESCARTAR?

En efecto:



CONFIRMAR => ALTA ESPECIFICIDAD

DESCARTAR => ALTA SENSIBILIDAD

EXACTITUD DE UNA PRUEBA DIAGNÓSTICA

La **exactitud** se define por la **sensibilidad**, la **especificidad**, los **valores predictivos** y la **eficacia**.

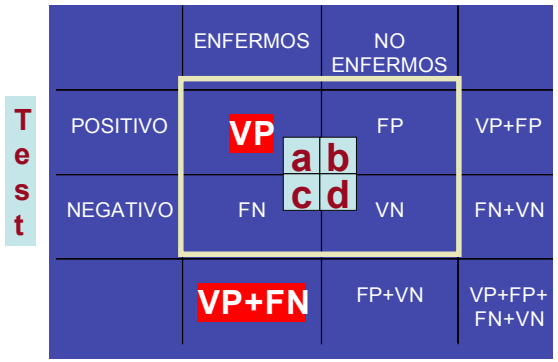
EFICACIA: Porcentaje de individuos correctamente clasificados.

$$\text{Eficacia} = \frac{VP + VN}{VP + VN + FP + FN}$$

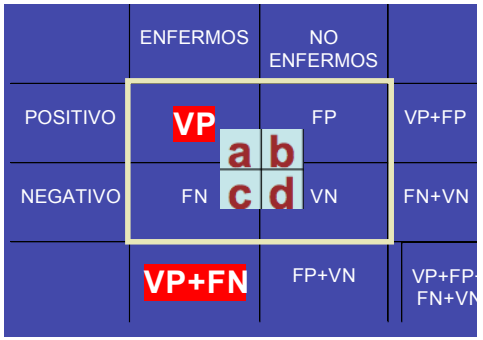
VP=Verdaderos positivos FP=Falsos posistivos
VN=Verdaderos negativos FN=Falsos negativos

¿Y si los datos vienen en una tabla tetracórica?

Prueba referencia

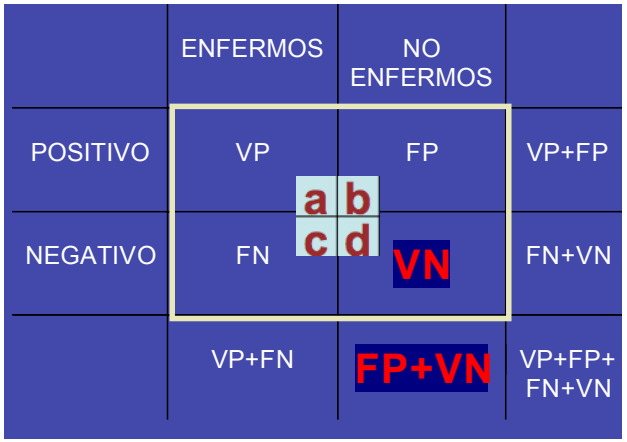


Sensibilidad



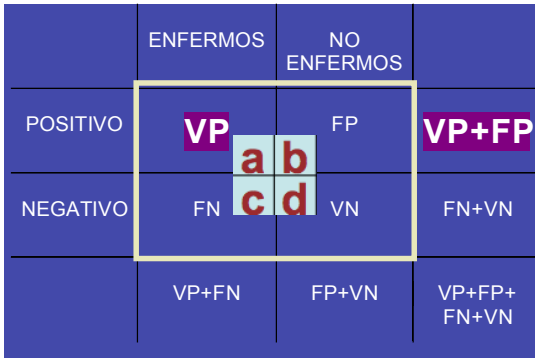
Sensibilidad=P(+/E) = a/a+c

Especificidad



Especificidad=P(-/noE)= d/d+b

Valor predictivo positivo (VPP):



P(E/+):VP/(VP+FP)= **a/a+b**

Valor predictivo negativo (VPN):

	ENFERMOS	NO ENFERMOS	
POSITIVO	VP (a)	FP (b)	VP+FP
NEGATIVO	FN (c)	VN (d)	FN+VN
	VP+FN	FP+VN	VP+FP+ FN+VN

P(noE/-): VN/(FN+VN)= **d/c+d**

Prueba referencia

Test	ENFERMOS	NO ENFERMOS	
POSITIVO	VP (a)	FP (b)	VP+FP
NEGATIVO	FN (c)	VN (d)	FN+VN
	VP+FN	FP+VN	VP+FP+ FN+VN

$$VPP = \frac{prevalencia \,.\, sensibilidad}{prevalencia \,.\, sensibilidad + (1- prev) \,.\, (1 - espec.)}$$

$$Prevalencia: \frac{a+c}{N}; Sensib.: \frac{a}{a+c}; Especif.: \frac{d}{b+d}$$

VPP

$$VPP = \frac{\frac{a+c}{N}\cdot\frac{a}{a+c}}{\frac{a+c}{N}\cdot\frac{a}{a+c}+\left(1-\frac{a+c}{N}\right)\cdot\left(1-\frac{d}{b+d}\right)} =$$

$$= \frac{\frac{a}{N}}{\frac{a}{N}+\frac{N-(a+c)}{N}\cdot\frac{b+d-d}{b+d}} =$$

$$= \frac{\frac{a}{N}}{\frac{a}{N}+\frac{b+d}{N}\cdot\frac{b}{b+d}} = \frac{a}{a+b} = VPP$$

Ejemplo:

Faringitis aguda. Exploración y Cultivo.

Validar la **impresión clínica** de faringitis estreptocócica: **SI/NO,** comparándolo con el **cultivo faríngeo** para estreptococo betahemolítico del grupo A (**estándar de oro**): **AUSENTE/PRESENTE**

	Presente	Ausente	
Sí	27	35	62
No	10	77	87
	37	112	149

Sensibilidad: 27/37: 0.73: **73%**

Especificidad: 77/112: 0.69: **69%**

Falsos Positivos: 35/112: **31.25%**

Falsos Negativos: 10/37: **27.02%**

Valor predictivo positivo: 27/ 62= 0.44: **44%**

Valor predictivo negativo: 77/ 87= 0.88: **88%**

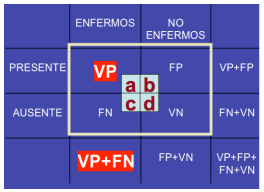
Eficacia: (27+77)/149: 0.70: 70%

Más información en una tabla tetracórica

Enfermedad

Factor Riesgo

	ENFERMOS	NO ENFERMOS	
PRESENTE	VP (a)	FP (b)	VP+FP
AUSENTE	FN (c)	VN (d)	FN+VN
	VP+FN	FP+VN	VP+FP+ FN+VN



Riesgo Relativo

$$RR = \frac{P(E/+)}{P(E/-)} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

La enfermedad es RR más frecuente entre los casos expuestos al factor de riesgo que en aquellos que no lo están.

Riesgo Relativo

El **RR no puede ser utilizado en estudios retrospectivos** ya que no se conocen las probabilidades condicionadas de presentar la enfermedad.

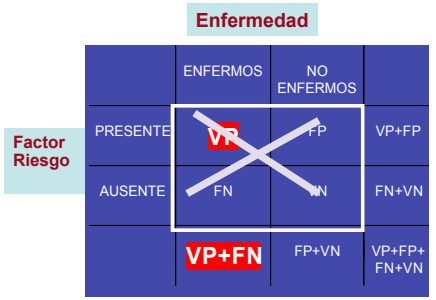
Fijamos de entrada los casos totales con enfermedad y los casos totales sin enfermedad y no es posible conocer la proporción real en los 2 subgrupos estudiados.

Fracción Atribuible

- Estima la proporción de la enfermedad, entre los expuestos, que puede ser atribuible al hecho de estar expuestos.
- La fracción atribuible en el grupo expuesto (**fracción etiológica**, o **porcentaje de riesgo atribuible en los expuestos**), establece el grado de influencia que tiene la exposición en la presencia de enfermedad entre los expuestos.
- Su cálculo se realiza: $Fa_{\text{en expuestos}} = (RR-1/RR)$

ODDs Ratio :PROSPECTIVOS

$$OR = \frac{\text{razón enfermos en los expuestos}}{\text{razón enfermos en los no expuestos}}$$



$$OR = \frac{\frac{P(E/+)}{P(noE/+)}}{\frac{P(E/-)}{P(noE/-)}} = \frac{\frac{a/a+b}{b/a+b}}{\frac{c/c+d}{d/c+d}} = \frac{a.d}{b.c}$$

Se fija el número de individuos expuestos al riesgo y el número de no expuestos.

Observamos cómo se desarrolla la enfermedad.

RETROSPECTIVOS:

$$OR = \frac{\text{razón de expuestos al riesgo en enfermos}}{\text{razón de expuestos al riesgo en sanos}}$$

Enfermedad

Factor Riesgo	ENFERMOS	NO ENFERMOS	
PRESENTE	VP	FP	VP+FP
AUSENTE	FN	VN	FN+VN
	VP+FN	FP+VN	VP+FP+ FN+VN

Se fija el número de individuos con la enfermedad y el número de individuos sin ella.

Observamos en cuantos está presente el síntoma.

$$OR = \frac{\frac{P(+/E)}{P(-/E)}}{\frac{P(+/noE)}{P(-/noE)}} = \frac{\frac{a/a+c}{c/a+c}}{\frac{b/b+d}{d/b+d}} = \frac{a.d}{b.c}$$

Un ejemplo

Tabla 5
Factores de riesgo de Sibilancias, Odds Ratio (OR) con Intervalo de Confianza (IC) al 95%, SiLBIDOS (Var 14)

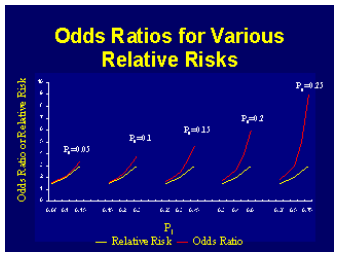
	n	%	O R	IC del 95% OR
Sexo Varón/Mujer	200	54.2	1.15	0.69 ; 1.47
PRN				
PRN 2500 a 3499 gr			1	Referencia
PRN ≤ 2499 gr	51	13,8	0.87	0.60 ; 1.24
PRN> 3500 gr	104	28.2	1.26	0.94 ; 1.68

Mascotas en casa al nacer el niño				
No mascotas			1	Referencia
Perro	38	10.3	0.97	0.65 ; 1.46
Gato	7	1.9	0.55	0.24 ; 1.27
Aves	21	5.7	1.31	0.75 ; 2.28
Conejo/hamster	12	3.3	**4.42**	**1.65; 11.90**
Otras	9	2.4	2.18	0.86 : 5.55

OR versus RR

Cuando los riesgos en ambos grupos son pequeños, el *odds ratio* se aproxima bastante al *riesgo relativo* y puede considerarse como una buena aproximación de éste.

Cuando se trata de eventos frecuentes, la discrepancia entre ambos parámetros se acentúa.



Cuando el evento es frecuente, RR y OR no son intercambiables.

La probabilidad de enfermar = (OR/OR+1)

Si el OR fuese, por ejemplo, de 2,5 entonces, aplicando la fórmula, podemos afirmar que la probabilidad de enfermar es de 0,714, es decir la probabilidad de que enferme un expuesto es de 71,4%.

Si el OR = 1, la probabilidad es del 50%, es decir que existen en este último caso la misma probabilidad de que el evento ocurra estando o no la otra variable en estudio presente.

Purificación GALINDO VILLARDON
pgalindo@usal.es

Departamento de ESTADÍSTICA
Universidad de Salamanca

ANÁLISIS RELACIÓN 2 VARIABLES CUANTITATIVAS:

Coeficiente de Correlación de Pearson

- **HIPOTESIS DE PARTIDA: H_0**
  Las dos variables en estudio son independientes

- **HIPOTESIS ALTERNATIVA: Ha**
  Las dos variables en estudio están relacionadas

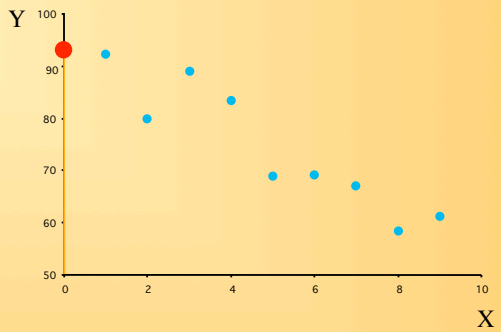
¿CÓMO NOS DECIDIMOS POR UNA U OTRA HIPÓTESIS?

- **Se recogen datos y se inspeccionan**

  - Gráficamente
  - Analíticamente

DIAGRAMA DE DISPERSIÓN

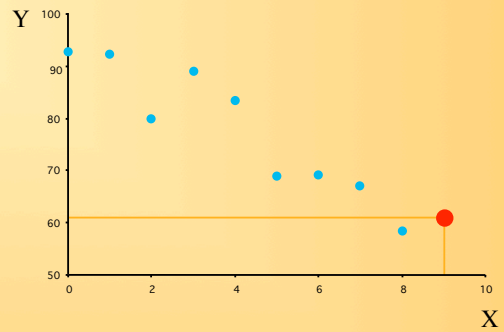
X	Y
0	92,8
1	92,3
2	80,0
3	89,1
4	83,5
5	68,9
6	69,2
7	67,1
8	58,3
9	61.2



Cada individuo vendrá representado por un punto (x_i,y_i) en el gráfico

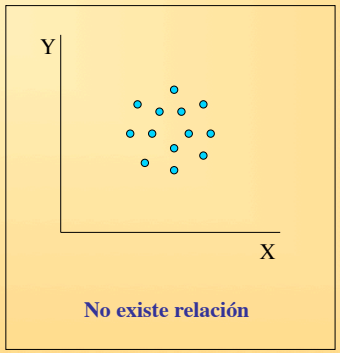
DIAGRAMA DE DISPERSIÓN

X	Y
0	92,8
1	92,3
2	80,0
3	89,1
4	83,5
5	68,9
6	69,2
7	67,1
8	58,3
9	61.2



Cada individuo vendrá representado por un punto (x_i,y_i) en el gráfico

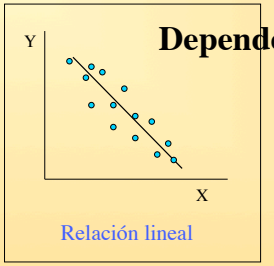
VARIABLES INDEPENDIENTES



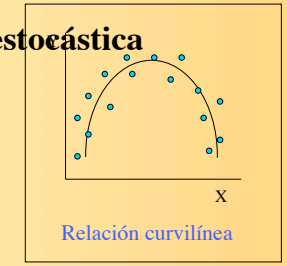
No existe relación

VARIABLES DEPENDIENTES

Dependencia estocástica

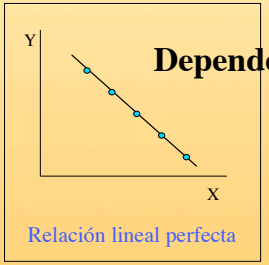


Relación lineal

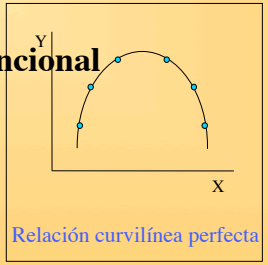


Relación curvilínea

Dependencia funcional



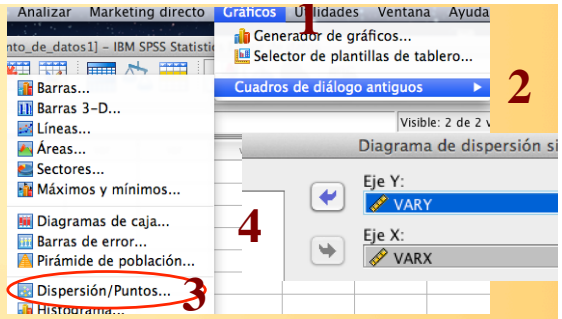
Relación lineal perfecta

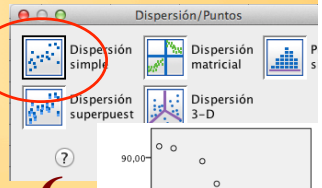


Relación curvilínea perfecta

DATOS

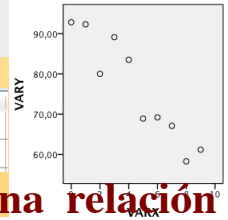
VARX	VARY
0	92,80
1	92,30
2	80,00
3	89,10
4	83,50
5	68,90
6	69,20
7	67,10
8	58,30
9	61,20





Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
VARX	Numérico	8	0		Ninguna	Ninguna	8	Derecha	Escala
VARY	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala

6



Existe una relación

¿Cómo cuantificar esa relación?

COVARIANZA

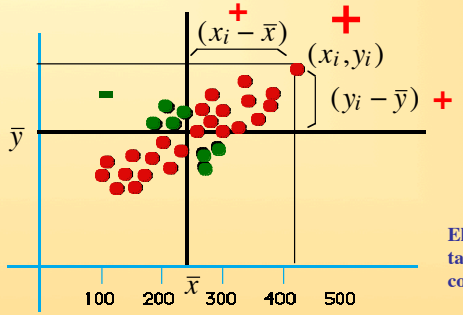
$$S_{xy} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

No está acotada

Arrastra las unidades de las dos variables

COVARIANZA Medida que muestra la relación entre dos variables cuantitativas.

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$



Seleccionamos un punto cuaquiera, por ejemplo en el primer cuadrante

La diferencia entre el valor de la coordenada *xi* y la media es positiva

La diferencia entre el valor de la coordenada *yi* y la media es positiva

El producto de ambos es positivo y por lo tanto lo es el correspondiente sumando de la covarianza

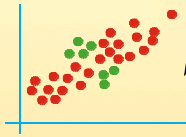
Todos los puntos del *primer cuadrante* aportan sumandos positivos (los pintados en rojo)

Para el *segundo cuadrante* la diferencia en x es positiva y en y negativa, por tanto el producto es negativo

Para el *tercer cuadrante*, ambas diferencias son negativas, por tanto el producto es positivo

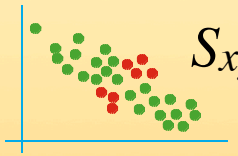
Para el *cuarto cuadrante*, el producto es negativo

COVARIANZA



$S_{xy} > 0$

Si la relación es directa la mayoría de los puntos aporta sumandos positivos y la COVARIANZA ES POSITIVA



$S_{xy} < 0$

Si la relación es inversa la mayoría de los puntos aporta sumandos negativos y la COVARIANZA ES NEGATIVA

$S_{xy} \approx 0$

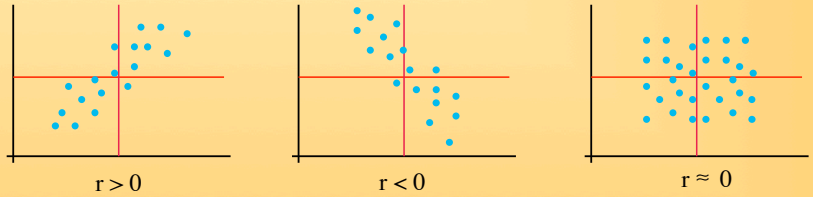
Si no hay relación se compensan los sumandos positivos y los negativos y la COVARIANZA ES APROXIMADAMENTE CERO.

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

$$r = \frac{s_{xy}}{s_x s_y}$$

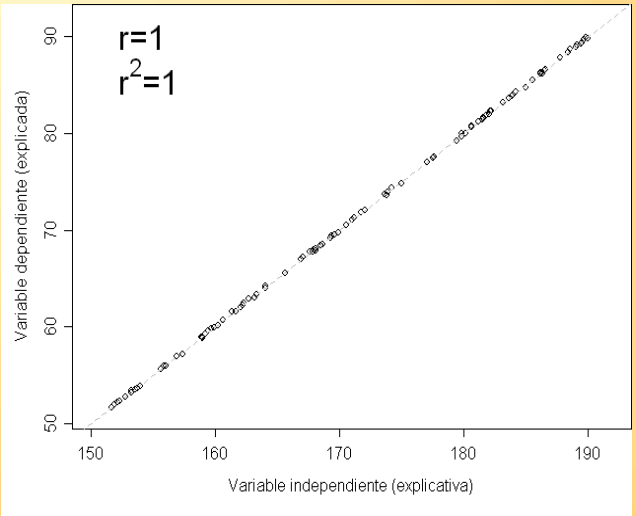
- $r < 0$ → Relación lineal inversa
- $r > 0$ → Relación lineal directa
- $r = 0$ → Variables independientes / Relación no lineal

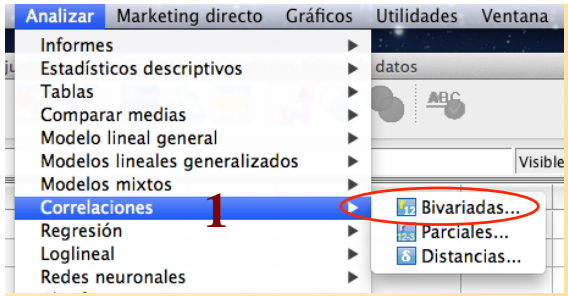
???

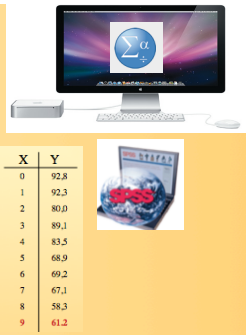


RELACIÓN LINEAL DIRECTA | RELACIÓN LINEAL INVERSA | INDEPENDIENTES

Evolución de r y diagrama de dispersión

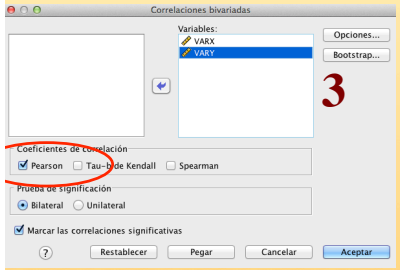






X	Y
0	92,8
1	92,3
2	80,0
3	89,1
4	83,5
5	68,9
6	69,2
7	67,1
8	58,3
9	61,2

2



Relación inversa

4

Correlaciones

		VARX	VARY
VARX	Correlación de Pearson	1	-,940**
	Sig. (bilateral)		,000
	N	10	10
VARY	Correlación de Pearson	-,940**	1
	Sig. (bilateral)	,000	
	N	10	10

**. La correlación es significativa al nivel 0,01 (bilateral).

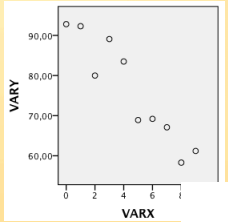
Sintetizando...

X	Y
0	92,8
1	92,3
2	80,0
3	89,1
4	83,5
5	68,9
6	69,2
7	67,1
8	58,3
9	61,2

1

- **HIPOTESIS DE PARTIDA: H_0**
  Las dos variables en estudio son independientes
- **HIPOTESIS ALTERNATIVA: Ha**
  Las dos variables en estudio están relacionadas

2



3

Relación inversa

4

Correlaciones

		VARX	VARY
VARX	Correlación de Pearson	1	-,940**
	Sig. (bilateral)		,000
	N	10	10
VARY	Correlación de Pearson	-,940**	1
	Sig. (bilateral)	,000	
	N	10	10

**. La correlación es significativa al nivel 0,01 (bilateral).

Purificación GALINDO VILLARDON
pgalindo@usal.es

Departamento de ESTADÍSTICA
Universidad de Salamanca

REGRESIÓN SIMPLE

Coeficiente de Regresión

X e Y variables cuantitativas

X: variable independiente

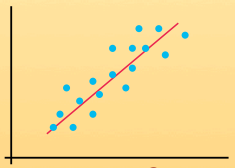
Y: variable dependiente

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

$$r = \frac{s_{xy}}{s_x s_y}$$

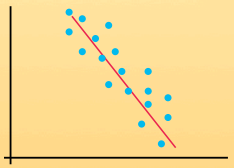
- $r < 0$ → Relación lineal inversa
- $r > 0$ → Relación lineal directa

RELACIÓN LINEAL DIRECTA



$r > 0$

RELACIÓN LINEAL INVERSA



$r < 0$

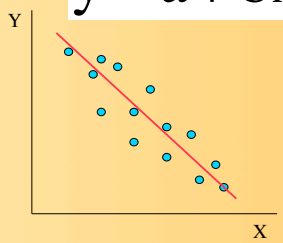
Si el p-valor es < 0.05 existe relación lineal significativa

ANÁLISIS DE REGRESIÓN

X: variable independiente

Y: variable dependiente

$$y = a + bx$$

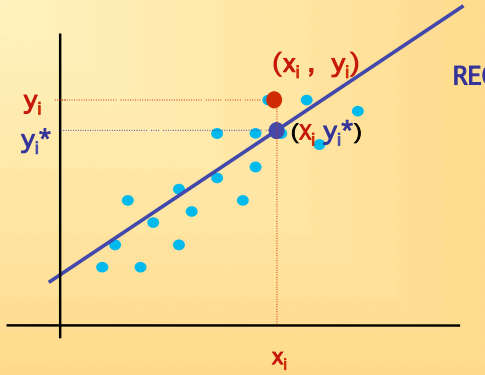


Existe regresión de los valores de una variable con respecto a los de otra, cuando hay una línea, llamada **línea de regresión**, que se ajusta a la nube de puntos.

A la ecuación que nos describe la relación entre las variables se le denomina **ecuación de regresión**.

CRITERIO DE LOS MÍNIMOS CUADRADOS

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - y_i^*\right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2$$



RECTA AJUSTADA

$y_i^* = a + b\ x_i$

CRITERIO DE LOS MÍNIMOS CUADRADOS

¿Cómo determinar los valores de a y b ?

Se hallan las derivas parciales de D respecto de **a** y **b**, y se resuelve el sistema resultante de igualar a 0 (minimizar) las ecuaciones obtenidas.

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - y_i^*\right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2$$

$$b = \frac{S_{xy}}{S_x^2}$$

Coefic. de Regresión

$$a = \bar{y} - b\bar{x}$$

Ordenada en el origen

COEFICIENTE DE REGRESIÓN

X: variable independiente

Y: variable dependiente

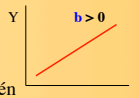
$Y = a + bX$

a: término independiente u ordenada en el origen

b: pendiente de la recta o **coeficiente de regresión**

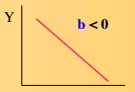
b > 0: **relación directa**

Cuando **X** aumenta **Y** también lo hace.



b < 0: **relación inversa**

Cuando **X** aumenta **Y** disminuye.



Coeficiente de Regresión

b > 0: incremento de **Y** cuando **X** aumenta en una unidad

b < 0: incremento de **Y** cuando **X** disminuye en una unidad

PODER EXPLICATIVO

COEFICIENTE DE DETERMINACIÓN: R^2

Poder explicativo / Bondad de Ajuste

$$0 \leq R^2 \leq 1$$

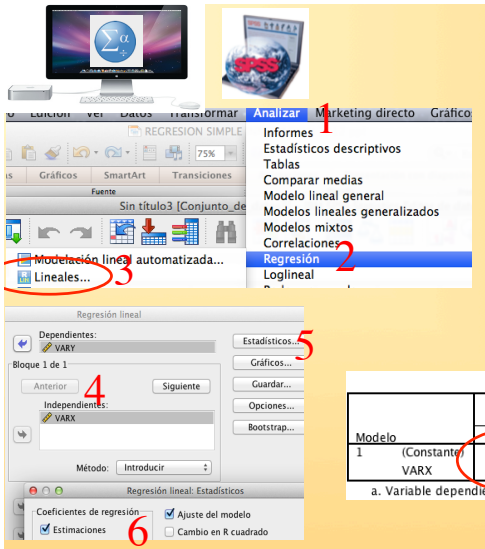
Cuanto más se aproxime R^2 a la 1, mayor **poder explicativo** o **mayor bondad de ajuste** del modelo.

$$R^2 = r^2$$

r = Coeficiente Correlación

R^2 x **100** = Porcentaje de variaciones explicadas por el modelo.

RESULTADOS



Modelo	R	R cuadrado
1	,940[a]	,884

R es el coeficiente de CORRELACIÓN

R^2 es el coeficiente de DETERMINACIÓN

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Error típ.	Coeficientes tipificados Beta	t	Sig.
1	(Constante)	a 94,273	2,746		34,333	,000
	VARX	b -4,007	,514	-,940	-7,791	,000

a. Variable dependiente: VARY

CONCLUSION

X e Y están relacionadas linealmente (p-valor =0.000) de manera indirecta. Cuando X aumenta una unidad, Y disminuye 4 unidades (en media)

$$Y=a+bX$$

P-valor

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Error típ.	Coeficientes tipificados Beta	t	Sig.
1	(Constante)	a 94,273	2,746		34,333	,000
	VARX	b -4,007	,514	-,940	-7,791	,000

a. Variable dependiente: VARY

Constante = a = Valor de la Y cuando X vale cero

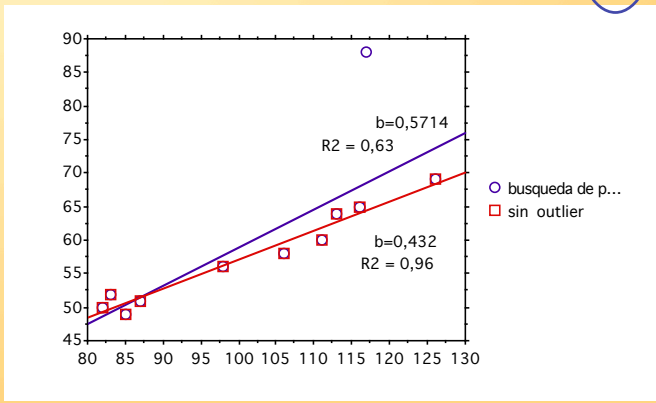
Variabilidad esperable para el coeficiente de regresión

Significa que a y b son significativamente distintos de cero y por tanto el modelo es de la forma

$$Y = 94.273 - 4.007$$

PRESENCIA DE OUTLIERS

82	83	85	87	98	106	111	113	116	**117**	126
50	52	49	51	56	58	60	64	65	**88**	69



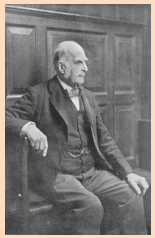
Purificación GALINDO VILLARDON
pgalindo@usal.es

Purificación VICENTE GALINDO
purivg@usal.es

Departamento de ESTADÍSTICA
Universidad de Salamanca

Regresión Simple

$$Y = a + bX$$



REGRESIÓN MÚLTIPLE

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

ESTIMADORES

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Los coeficientes de regresión :

Los coeficientes de regresión se interpretan como el efecto de la variable X_j en la variable dependiente Y, cuando el resto de las variables se mantienen constantes; es decir, el cambio producido en la variable dependiente Y por cada incremento unitario en la regresora X_j manteniendo constante el resto de las predictoras.

Coeficientes de regresión estandarizados:
Cada $Beta_i$ (estandarizado) se interpreta como el cambio, en unidades de desviación típica, en la variable dependiente, por cada cambio en una desviación típica en la variable independiente Xi, manteniendo el resto de las variables independientes constantes.

Equivale a realizar una regresión sobre datos estándarizados
(media cero, desviación típica 1)

Diabetes, colesterol y tratamiento de la hipertensión

Feher, M.D.; Rains, S.G.H.; Richmond, W.; Torrens, D. et al.

" Beta-blockers, Lipoproteins and Noninsulin-Dependent Diabetes"

Postgrad. Med. Journal. Vol 64: 926-930, 1988

REGRESION MULTIPLE

Se ha llevado a cabo un estudio sobre **71 individuos** varones en el que se pretende conocer el efecto de los Beta Bloqueantes sobre HDL (concretamente sobre la subfracción HDL-2) en los hipertensos diabéticos.

Para ello se consideran 8 variables explicativas

$$B = \begin{cases} 1 & \text{si toma } \beta-\text{bloqueadores} \\ 0 & \text{no los toma} \end{cases}$$

$$D = \begin{cases} 1 & \text{si bebe} \\ 0 & \end{cases}$$

$$S = \begin{cases} 1 & \text{si fuma} \\ 0 & \end{cases}$$

A : Edad

W : Peso

T : Trigliceridos

C : C-peptidos

G : Glucosa

HIPOTESIS: Tratar la hipertensión de los diabéticos con β – bloqueantes puede empeorar su perfil de colesterol.

Modelo: **H**= 0.711 - 0.0824·**B** - 0.0173·**D** - 0.0399·**S** - 0.00455·**A** - 0.00214·**W** - 0.0444·**T** + 0.00463·**C** - 0.00391·**G**

H ≡ HDL - 2
B ≡ ¿β – bloqueadores ?
D ≡ ¿Bebe?
S ≡ ¿Fuma?
A ≡ Edad
W ≡ Peso
T ≡ Trigliceridos
C ≡ C - Peptid.
G ≡ Glucosa

Analisis de la Varianza

Fuentes	g.l.	S.Cuad.	Med.Cuad.	F-exp	p-valor
Regresión	8	0.546959	0.068370	11.40	**0.000**
Error	62	0.371915	0.005999		
Total	70	0.918873			

Predictor	*Coef*	*Error. stand.*	*t-exp*	*p-valor*
Constant	0,7110	0,1102	6,45	0,000
B	- 0,08244	0,02293	- 3,59	**0,001**
D	-0,01726	0,02121	- 0,81	0,419
S	- 0,03995	0,02078	- 1,92	0,059
A	- 0,004549	0,001179	- 3,86	**0,000**
W	- 0,002140	0,002722	- 0,79	0,435
T	- 0,044372	0,009411	- 4,71	**0,000**
C	0,004633	0,007811	0,59	0,555
G	- 0,003907	0,003239	- 1,21	0,232

$R^2 = 59.5\%$ $\mathbf{R^2}$(ajustado) = **54.3%**

RESULTADOS

Las variables **D** (bebe), **W** (peso), **C** (C-peptidos) y **G** (nivel de glucosa) son **no significativas** (**P>0.05**) por lo tanto sus coeficientes no difieren significativamente de cero.

A partir de estos datos no se puede pensar que afecten los niveles de HDL-2.

El caso de la variable **S** (¿fuma?) es dudoso . **Los demás** coeficientes son estadísticamente **significativos** (**P<0.05**). Esto significa que no pueden considerarse nulos y por tanto que las variables correspondientes **B** (beta - bloqueadores), **A** (edad) y **T** (triglicéridos) afectan los niveles de HDL-2.

Analizando los signos correspondientes a las tres variables significativas (**P<0.05**) podemos afirmar que **B**, **A** y **T** afectan negativamente los niveles de HDL-2; es decir, cuanto más edad, cuanto más altos sean los triglicéridos, y cuanto mayor sea la dosis de beta-bloqueadores, el HDL-2 disminuye, en hipertensos diabéticos.

CONCLUSION

Los beta-bloqueadores bajan el HDL en hipertensos diabéticos, por lo que deben ser usados con mucha precaución en diabéticos, o ser evitados.

Regresión Logística

M Purificación Galindo Villardón

pgalindo@usal.es

Introducción

Herramienta para modelizar la relación entre una variable dicotómica de respuesta y una o más variables predictoras.

Es posible incluir regresores de tipo cualitativo, mediante la utilización de variables indicadoras, de manera análoga a como se hace en regresión lineal.

REGRESIÓN LOGÍSTICA

Variable respuesta dicotómica

Y=1 si el hecho ocurre
Y=0 si el hecho no ocurre

$$Y=\beta_0 + \beta_1 X_1$$

PROBLEMA:

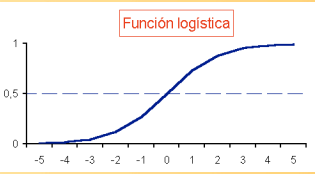
El valor estimado debe ser un valor entre cero y uno por ser una probabilidad

La regresión lineal debe ser descartada

REGRESIÓN LOGÍSTICA

~~$$Y=\beta_0 + \beta_1 X_1$$~~

$$P(Y=1)=\frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$$



Con una sencilla transformación (logit) puede convertirse en lineal

$$\log it(p) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \quad \Rightarrow \frac{p}{1-p} = e^{\beta_0+\beta_1 X}$$

$$\Rightarrow p = (1-p)e^{\beta_0+\beta_1 X} \quad \Rightarrow p = e^{\beta_0+\beta_1 X} - p e^{\beta_0+\beta_1 X}$$

$$\Rightarrow p + p e^{\beta_0+\beta_1 X} = e^{\beta_0+\beta_1 X} \quad \Rightarrow p\left(1+e^{\beta_0+\beta_1 X}\right) = e^{\beta_0+\beta_1 X}$$

$$\Rightarrow p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Invirtiendo ambos términos de la ecuación…

$$\frac{1}{p} = \frac{e^{\beta_0+\beta_1 X}+1}{e^{\beta_0+\beta_1 X}} \quad \Rightarrow \frac{1}{p} = 1 + \frac{1}{e^{\beta_0+\beta_1 X}}$$

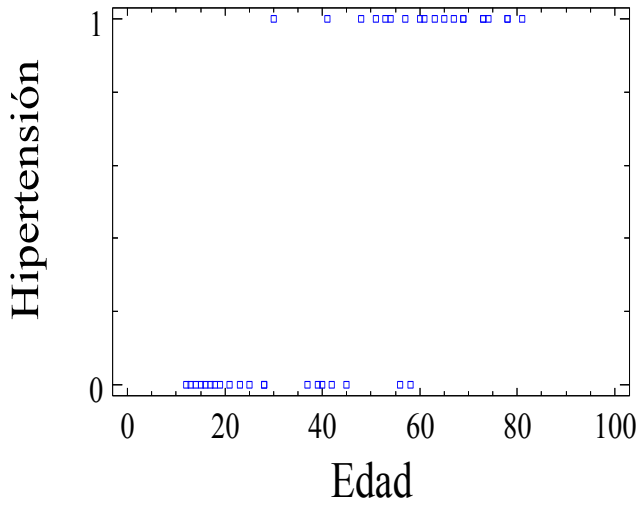
$$\Rightarrow \frac{1}{p} = 1 + e^{-(\beta_0+\beta_1 X)}$$

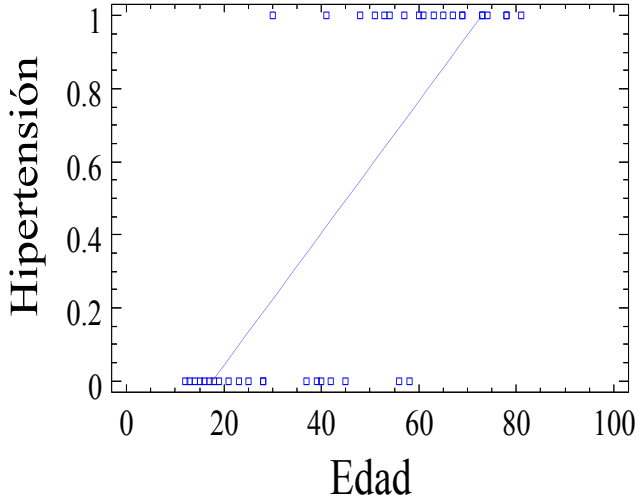
Invirtiendo nuevamente,

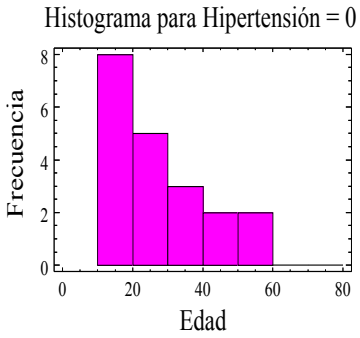
$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

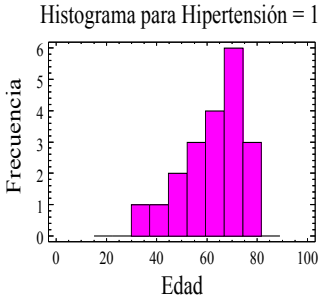
Al estimar β_0 y β_1 por Máxima Verosimilitud, se obtiene…

Edad	Hipertensión
69	Sí
23	No
53	Sí
58	No
78	Sí
39	No
67	Sí
37	No
65	Sí
41	Sí
69	Sí
30	Sí
14	No
73	Sí
21	No



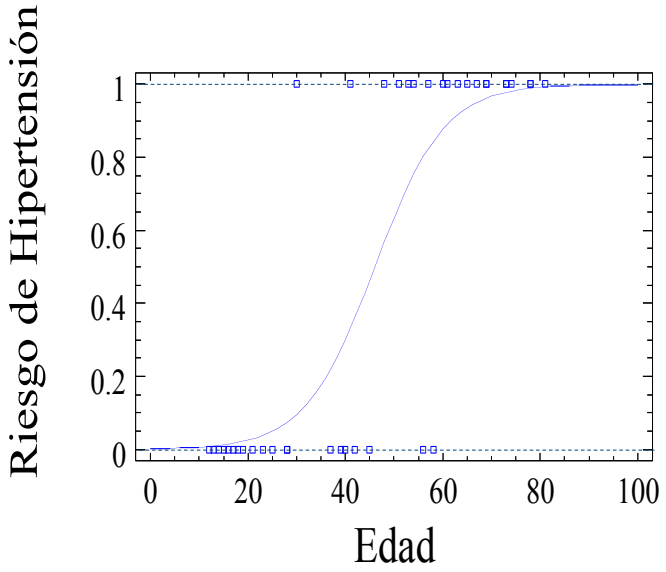






$$Odds = \frac{p}{(1-p)}$$

Límite Inferior	Límite Superior	Frecuencia de éxitos	Frecuencia de fracasos	Odds Ratio
10,0	25,4	0,00	0,55	0,00
24,4	40,8	0,05	0,25	0,20
40,8	56,2	0,25	0,15	1,67
56,2	71,6	0,40	0,05	8,00
71,6	87,0	0,30	0,00	Infinito



Estimación por Máxima Verosimilitud

Consiste en obtener los valores de los parámetros que maximizan la probabilidad de obtener la muestra observada.

En ocasiones es posible llegar a los estimadores usando herramientas del álgebra y el cálculo, pero en la mayoría de ocasiones se usan procesos numéricos iterativos.

Ensayando β_0= -6,48 β_1=0,14 en

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

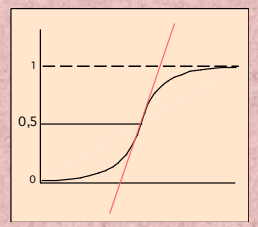
Edad	Resultado	Probabilidad estimada de éxito	Probabilidad estimada de fracaso
69	Sí	0,96	0,04
23	No	0,04	0,96
53	Sí	0,73	0,27
58	No	0,84	0,16
78	Sí	0,99	0,01
39	No	0,27	0,73
67	Sí	0,95	0,05
37	No	0,22	0,78
65	Sí	0,94	0,06
41	Sí	0,33	0,67
69	Sí	0,96	0,04
30	Sí	0,09	0,91
14	No	0,01	0,99
73	Sí	0,98	0,02
21	No	0,03	0,97

- Falso Positivo: Un sujeto es declarado como enfermo sin estarlo.
- Falso Negativo: Un sujeto se declara sano, estando realmente enfermo.

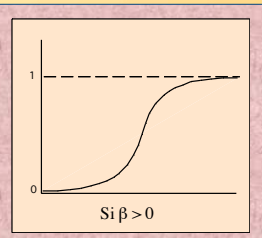
INTERPRETACIÓN DE LOS PARÁMETROS

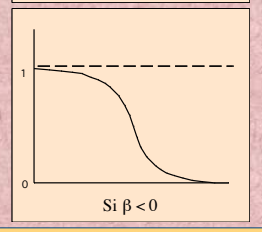
Regresión logística simple

β_1 es la pendiente de la recta tangente a la curva en el punto para el que la probabilidad esperada es 0.5



El valor de X para el que se alcanza la probabilidad 0,5 es $-\beta_0 / \beta_1$.





REGRESIÓN LOGÍSTICA

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$$

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots)}}$$

Función logística

Parámetros

REGRESIÓN LOGÍSTICA

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots)}}$$

X_1 , X_2 , X_3 ...
Pueden ser:
Dicotómicas, nominales, ordinales, o continuas

Individuos mayores de 40 años

Y =Desarrollo enfermedad coronaria en 10 años de seguimiento

$$P(Y=1)=\frac{1}{1+e^{-(-6.61+0.075X_1+0.312X_2+0.018X_3)}}$$

X_1 ,	X_2 ,	X_3
Edad	Hábito fumar	Tensión arterial
Continua	Dicotómica	Continua

Individuos mayores de 40 años

Y =Desarrollo enfermedad coronaria (EC) en 10 años de seguimiento

$$P(Y=1)=\frac{1}{1+e^{-(-6.614+0.075X_1+0.312X_2+0.018X_3)}}= 0.68$$

Se estima que:
El 68% de los sujetos con ese perfil desarrollarán un EC en el siguiente decenio

X_1 ,	X_2 ,	X_3
Edad	Hábito fumar	Tensión Sistólica
58 años	**Fumador**	**150**

Individuos mayores de 40 años

Y =Desarrollo enfermedad coronaria en 10 años de seguimiento

$$P(Y=1)=\frac{1}{1+e^{-(-6.61+0.075X_1+0.312X_2+0.018X_3)}}$$

$$(-6.61+0.075X_1+0.312X_2+0.018X_3)$$

Interpretación igual que en regresión múltiple

Interpretación de los **coeficientes de regresión** en términos de **Odds Ratio**

Odds y Probabilidad

$$Odd(E)=\frac{P(E)}{1-P(E)}$$

$$O(E)=\frac{p}{1-p}$$

Odd(E) = ?

P(E) = 0.68 => Odd(E)=0.68/(1-0.68)=2.12

Es 2.12 veces más probable enfermar que no enfermar

Odds Ratio

$$OR(E)=\frac{\frac{P_F(E)}{1-P_F(E)}}{\frac{P_{\bar{F}}(E)}{1-P_{\bar{F}}(E)}}$$

Odds Ratio

$$OR(E) = \frac{\frac{P_F(E)}{1-P_F(E)}}{\frac{P_{\bar{F}}(E)}{1-P_{\bar{F}}(E)}}$$

$P(E)_{\text{Fumador}} = 0.68$

$P(E)_{\text{No Fumador}} = 0.61$

$$OR(E) = \frac{\frac{0.68}{1-0.68}}{\frac{0.61}{1-0.61}} = \frac{2.125}{1.564} = 1.36$$

$$O(E_F) = e^{-(\beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_3 X_3)}$$

$$O(E_{\bar{F}}) = e^{-(\beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_3 X_3)}$$

$$O(E) = e^{\beta 2\,(1-0)} = e^{\beta 2} = 1.36$$

Interpretación de los parámetros

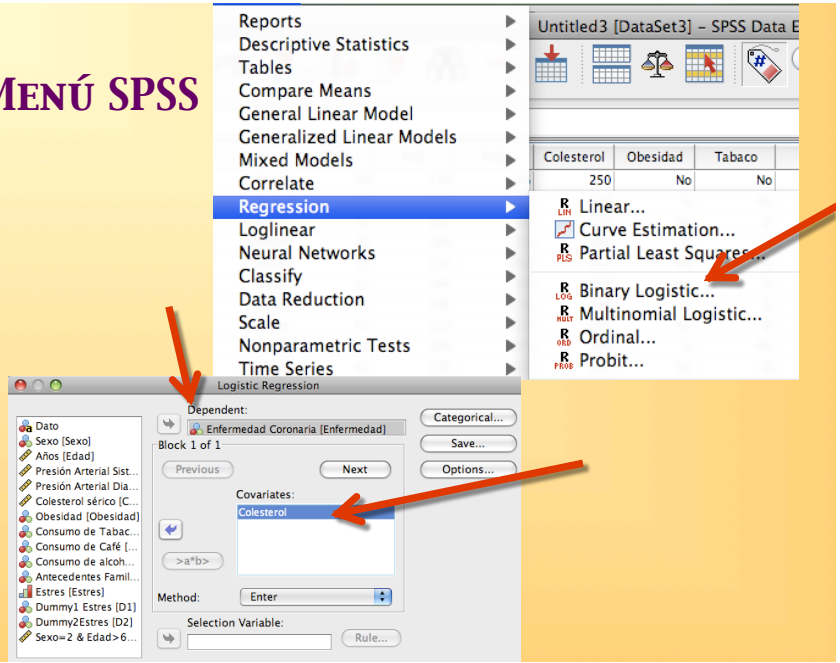
$$OR(E) = \frac{\frac{P_F(E)}{1-P_F(E)}}{\frac{P_{\bar{F}}(E)}{1-P_{\bar{F}}(E)}} = e^{\beta}$$

$$e^{\beta} = \frac{\frac{P(E=1/x+1)}{P(E=0/x+1)}}{\frac{P(E=1/x)}{P(E=0/x)}} = \text{Odds Ratio}$$

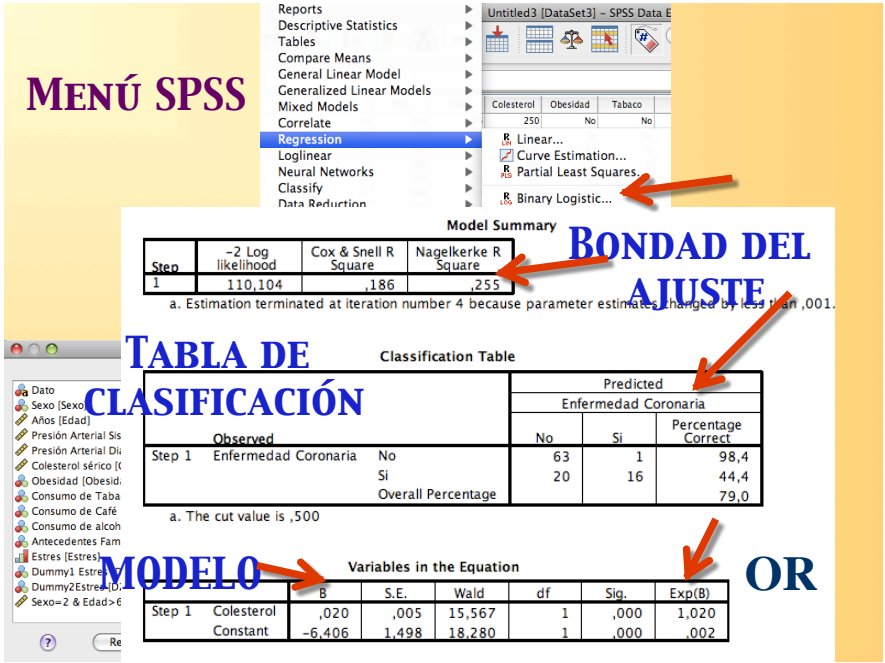
Para variables continuas:

$e^{(\beta 1)}$ mide el odds-ratio entre un individuo con un valor $x+1$ y un individuo con valor x en la variable independiente, para cualquier valor x

MENÚ SPSS



MENÚ SPSS



$$\text{Prob}\left[Enferm = SI \,/\, X = x_i\right] = p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

e^{β_1}

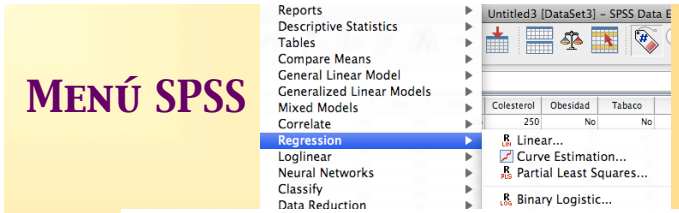
Estimadores | Errores estándar | Significaciones | Odds-Ratio

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	Colesterol (β_1)	,020	,005	15,567	1	,000	1,020
	Constant (β_0)	-6,406	1,498	18,280	1	,000	,002

Por cada incremento de una unidad en el **colesterol** el logit se incrementa en 0.02, por tanto el odds (ventaja) de "SI" estar enfermo frente a "NO" estar enfermo se multiplica por 1.020 al incrementarse una unidad el colesterol.

MENÚ SPSS



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	110,104	,186	,255

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

TABLA DE CLASIFICACIÓN

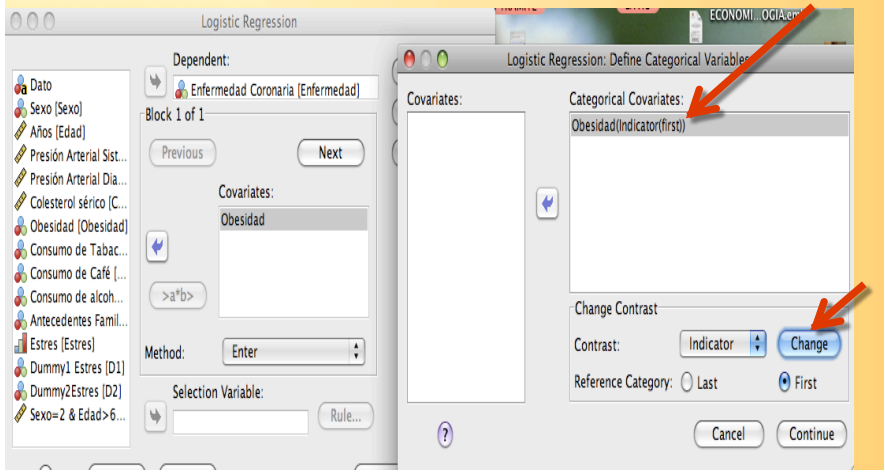
Classification Table

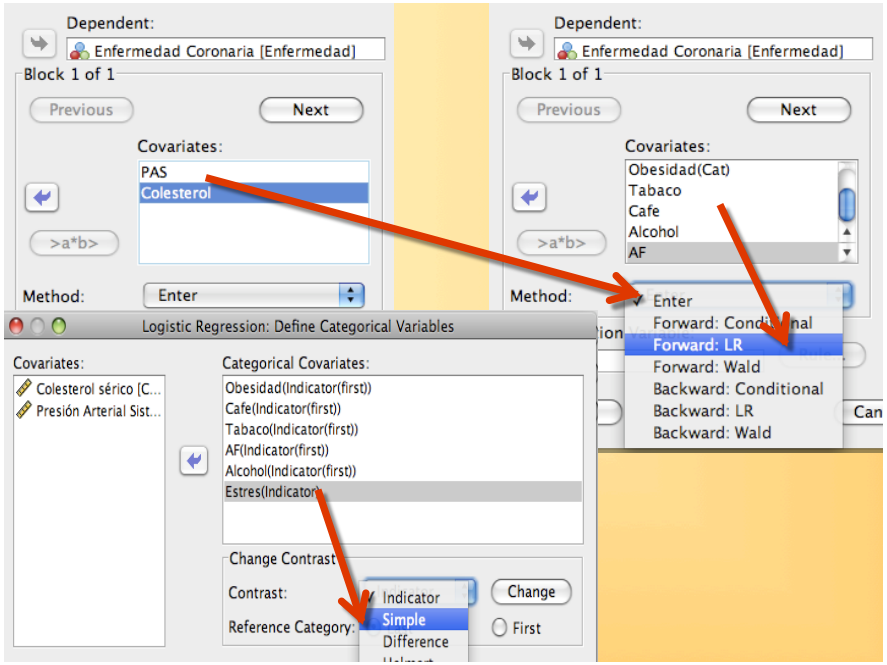
Observed			Predicted: Enfermedad Coronaria: No	Predicted: Enfermedad Coronaria: Si	Percentage Correct
Step 1	Enfermedad Coronaria	No	63	1	98,4
		Si	20	16	44,4
		Overall Percentage			79,0

El 98,4% de los que no tienen enfermedad coronaria se han clasificado correctamente utilizando el modelo de regresión logística, mientras que de los que "SI" tiene la enfermedad se clasifican correctamente el 44.4%. Conociendo los valores del Colesterol se predice correctamente la presencia/ausencia de enfermedad coronaria en el 79% de los sujetos.

Obsérvese que el modelo es mucho más adecuado para descartar que para confirmar.

Si es categórica…





Logistic Regression Variable Selection Methods

Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables.

Hide details

- **Enter.** A procedure for variable selection in which all variables in a block are entered in a single step.
- **Forward Selection (Conditional).** Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates.
- **Forward Selection (Likelihood Ratio).** Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates.
- **Forward Selection (Wald).** Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of the Wald statistic.
- **Backward Elimination (Conditional).** Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on conditional parameter estimates.
- **Backward Elimination (Likelihood Ratio).** Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on the maximum partial likelihood estimates.
- **Backward Elimination (Wald).** Backward stepwise selection. Removal testing is based on the probability of the Wald statistic.

Logistic Regression Define Categorical Variables

You can specify details of how the Logistic Regression procedure will handle categorical variables:

Covariates. Contains a list of all of the covariates specified in the main dialog box, either by themselves or as part of an interaction, in any layer. If some of these are string variables or are categorical, you can use them only as categorical covariates.

Categorical Covariates. Lists variables identified as categorical. Each variable includes a notation in parentheses indicating the contrast coding to be used. String variables (denoted by the symbol < following their names) are already present in the Categorical Covariates list. Select any other categorical covariates from the Covariates list and move them into the Categorical Covariates list.

Change Contrast. Allows you to change the contrast method. Available contrast methods are:

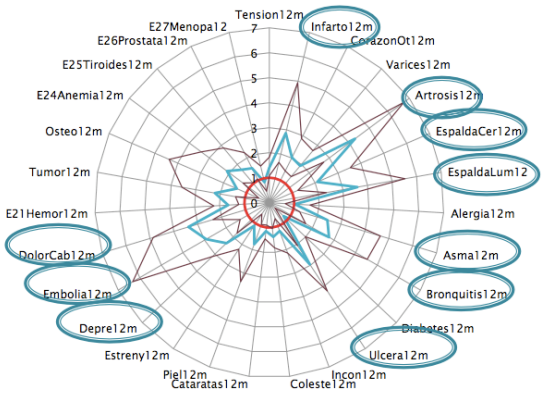
- **Indicator.** Contrasts indicate the presence or absence of category membership. The reference category is represented in the contrast matrix as a row of zeros.
- **Simple.** Each category of the predictor variable (except the reference category) is compared to the reference category.
- **Difference.** Each category of the predictor variable except the first category is compared to the average effect of previous categories. Also known as reverse Helmert contrasts.
- **Helmert.** Each category of the predictor variable except the last category is compared to the average effect of subsequent categories.
- **Repeated.** Each category of the predictor variable except the first category is compared to the category that precedes it.
- **Polynomial.** Orthogonal polynomial contrasts. Categories are assumed to be equally spaced. Polynomial contrasts are available for numeric variables only.
- **Deviation.** Each category of the predictor variable except the reference category is compared to the overall effect.

If you select Deviation, Simple, or Indicator, select either First or Last as the reference category. Note that the method is not actually changed until you click Change.

String covariates must be categorical covariates. To remove a string variable from the Categorical Covariates list, you must remove all terms containing the variable from the Covariates list in the main dialog box.

ENCUESTA NACIONAL DE SALUD
ENS2006
Fichero Adultos
65-80 años

RESULTADOS



BIBLIOGRAFIA

Luis Carlos Silva Ayçaguer
Excursión a la regresión logística en ciencias de la salud.
Ed. Díaz de Santos Madrid 1995

David W. Hosmer y Stanley Lemeshow
Applied Logistic Regression
Ed. John Wiley New York 1989

ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)
ANÁLISIS FACTORIAL

Dra. Purificación Galindo
pgalindo@usal.es

ACP

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables).

Es decir, ante un conjunto de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible.

Los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

ACP

Variabilidad
es equivalente a
Información

EJEMPLOS TIPICOS

-Las diferentes asignaturas que componen la enseñanza media se dividen en Ciencias y Letras.

-Ciertos síntomas clínicos propios de los enfermos mentales se clasifican en síntomas de tipo neurótico y síntomas de tipo psicótico.

-El estudio de los conflictos internos de las naciones descubre la existencia de tres factores: agitación, revolución y subversión.

-Los ítems de un test de BURNOUT conforman tres dimensiones latentes: Autoestima, Agotamiento y Despersonalización

ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

Pearson, K. (1901). On lines and planes of closets fit to systems of points in the space. *Philosophical Magazine*, **2**: 559-572.

Pearson trata de encontrar una matriz de menor dimensión que la original, que mejor resuma la información de los datos originales, en el sentido de los mínimos cuadrados.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**:417-441,498-520.

Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika*, **1**: 27-35

La aproximación de Hotelling obtiene sucesivamente combinaciones lineales de variables con varianza máxima.

¿Cómo se calculan las Componentes Principales?

Pearson y Hotteling demuestran que:

Basta con:

1) Calcular la matriz de covarianzas (o correlaciones)

Sus varianzas

$\mathbf{S_{(pxp)}=X'X}$:

Componentes Principales

2) Buscar los **valores propios** y los **vectores propios** de esa matriz de covarianzas (o correlaciones)

ALGUNAS PROPIEDADES DE LAS COMPONENTES

PROPORCIÓN DE VARIANZA ABSORBIDA POR CADA COMPONENTE

$$\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} x100$$

El valor propio asociado a cada variable latente, dividido por la suma de todos ellos, nos indica la importancia relativa de la correspondiente variable latente

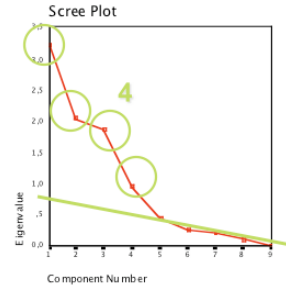
¿Cuántos factores debemos retener?

VALORES PROPIOS MAYORES QUE UNO

El número de factores estará determinado por el número de valores prop

REGLA DEL 75% DE LA VARIANZA

El número de factores esta determinado por la absorción de inercia. Se t como sean necesarios para conseguir un 75% de inercia absorbida.



REGLA DEL CODO (CATTELL (1966) SCREE PLOT

El procedimiento del scree plot de Cattell consiste en representar gráficamente los valores propios en orden descendente y dibujar una recta a través de las componentes con los valores propios más bajos. Se retienen las componentes que se corresponden con los autovalores que quedan por encima de la línea.

ESPECIFICACION DEL USUARIO

Es posible especificar personalmente el número de factores. Usualmente no será mayor que el número de variables dividido por dos. En el caso en que se sobrestime éste valor, el nº será ajustado por el ordenador.

¿QUÉ HACER PARA INTERPRETAR LOS EJES FACTORIALES?

- Se analizan las saturaciones (en valor absoluto). Aquellas variables que presentan altas saturaciones son las que tiene mayor importancia en la interpretación del eje.

- Las más interesantes suelen ser las que presentan altas saturaciones para ese eje y bajas para los demás

EJEMPLO ACP

X_{20x7}

n=20 pacientes
p=7 variables

X_1=Presión arterial media (mmHg)
X_2=Edad (años)
X_3=Peso (kg.)
X_4=Superficie corporal (m^2)
X_5=Duración de la hipertensión (años)
X_6=Pulso (pulsaciones/minuto)
X_7=Medida del estress (0-100)

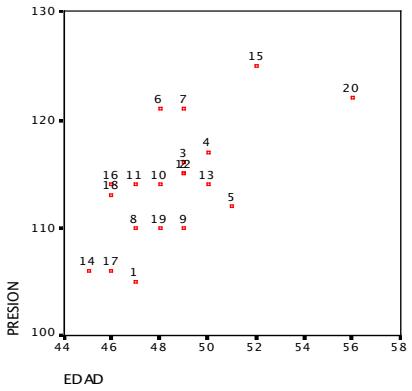
X_1	X_2	X_3	X_4	X_5	X_6	X_7
105	47	85,4	1,75	5,1	63	33
115	49	94,2	2,1	3,8	70	14
116	49	95,3	1,98	8,2	72	10
117	50	94,7	2,01	5,8	73	99
112	51	89,4	1,89	7	72	95
121	48	99,5	2,25	9,3	71	10
121	49	99,8	2,25	2,5	69	42
110	47	90,9	1,9	6,2	66	8
110	49	89,2	1,83	7,1	69	62
114	48	92,7	2,07	5,6	64	35
114	47	94,4	2,07	5,3	74	90
115	49	94,1	1,98	5,6	71	21
114	50	91,6	2,05	10,2	68	47
106	45	87,1	1,92	5,6	67	80
125	52	101,3	2,19	10	76	98
114	46	94,5	1,98	7,4	69	95
106	46	87	1,87	3,6	62	18
113	46	94,5	1,9	4,3	70	12
110	48	90,5	1,88	9	71	99
122	56	95,7	2,09	7	75	99

X_{20x7}

105	47	85,4	1,75	5,1	63	33
115	49	94,2	2,1	3,8	70	14
116	49	95,3	1,98	8,2	72	10
117	50	94,7	2,01	5,8	73	99
112	51	89,4	1,89	7	72	95
121	48	99,5	2,25	9,3	71	10
121	49	99,8	2,25	2,5	69	42
110	47	90,9	1,9	6,2	66	8
110	49	89,2	1,83	7,1	69	62
114	48	92,7	2,07	5,6	64	35
114	47	94,4	2,07	5,3	74	90
115	49	94,1	1,98	5,6	71	21
114	50	91,6	2,05	10,2	68	47
106	45	87,1	1,92	5,6	67	80
125	52	101,3	2,19	10	76	98
114	46	94,5	1,98	7,4	69	95
106	46	87	1,87	3,6	62	18
113	46	94,5	1,9	4,3	70	12
110	48	90,5	1,88	9	71	99
122	56	95,7	2,09	7	75	99

Considerando sólo 2 variables

Podemos interpretar la similitud entre individuos



X_{20x7}

Si considerásemos las **7 variables**,

Necesitaríamos un **hiperespacio** de 7 dimensiones para representar a los sujetos.

Podemos simplificar el problema calculando las Componentes Principales (2 por ejemplo)

ACP

$S_{(pxp)}=X'X$

En nuestro ejemplo

•Los valores propios (λ_j) representan la varianza de las nuevas variables; es decir su capacidad informativa

λ_j	3,908	1,470	0,708	0,521	0,308	0,080	0,002	7
% Var.	55,832	21,003	10,125	7,452	4,399	1,154	0,032	100%
% Var. acum.	55,382	76,835	86,961	94,414	98,813	99,968	100	

$S_{(pxp)}=X'X$

•Los vectores propios son las componentes principales

En nuestro ejemplo

Factores de carga

X_i	Y_1	Y_2	Y_3
PRESION	0,48814	-0,18969	-0,00547
EDAD	0,36568	0,25049	-0,15331
PESO	0,44713	-0,33244	0,03614
SUPERFICIE	0,40671	-0,38985	0,00711
DURACION	0,21965	0,43261	0,86381
PULSO	0,42683	0,23457	-0,16222
STRESS	0,17952	0,62976	-0,45015

El resto aporta muy poca información

1ª Componente

Y_1 = 0,48 PRESIÓN + 0,36 EDAD + … + 0,17 STRESS

SALIDA DE SPSS

Component Matrix[a]

	Component 1	Component 2
PRESION	,965	-,230
PESO	,884	-,403
PULSO	,844	,284
SUPERFIC	,804	-,473
EDAD	,723	,304
ESTRÉS	,355	,764
DURACIÓN	,434	,525

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

Correlaciones entre las componentes principales y las variables observadas

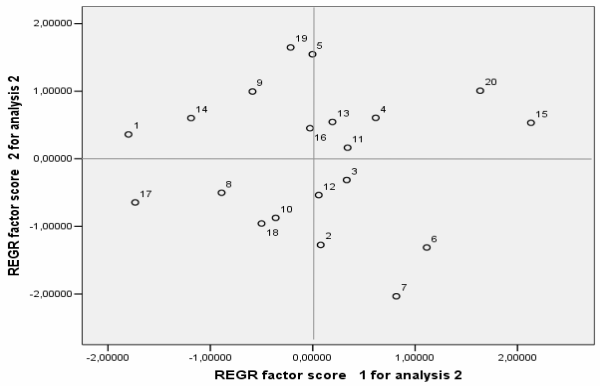
Interpretación de las variables latentes

1: Eje horizontal

ÍNDICE DE RIESGO DE ENFERMEDAD CORONARIA

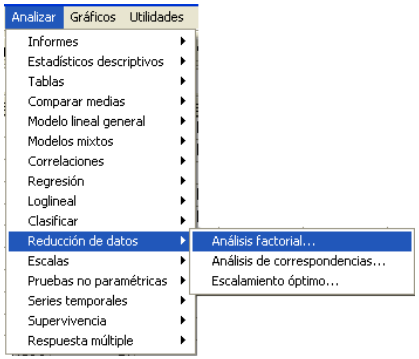
2: Eje vertical

ESTRÉS



- No existe en el programa SPSS una opción propia para realizar un ACP.

¿Cómo realizar un *PCA* en SPSS?



Para hacerlo, deberemos recurrir a la opción Análisis Factorial (AF)

ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)
ANÁLISIS FACTORIAL (AF)

ACP ***AF***

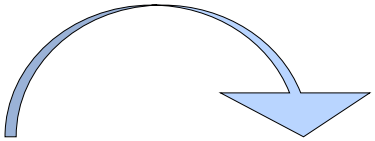
Con ambas técnicas, tratamos de reducir la dimensionalidad de los datos

*Tratamos de recoger la información de las **p variables observables** con unas pocas variables (q<p), no directamente observables (latentes), que recojan un alto porcentaje de la información original, y que sean incorreladas*

Situación de partida

p variables

con mucha variabilidad y altamente correlacionadas



ACP ***AF***

Buscamos

q variables

(q<p)

con mucha variabilidad pero independientes

ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

ACP

AF

*Estudiamos la estructura de correlaciones entre VARIABLES

*Se buscan variables hipotéticas que EXPLIQUEN las variables originales

*Representamos CORRELACIONES entre variables y entre variables y factores

ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

ACP

*Nos interesa la información de los INDIVIDUOS

*Queremos describir los valores de los individuos mediante un pequeño nº de variables, que sean combinación de las originales

*Representamos INDIVIDUOS

AF

*Estudiamos la estructura de correlaciones entre VARIABLES

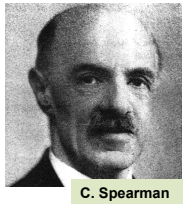
*Se buscan factores hipotéticos que EXPLIQUEN las variables originales

*Representamos CORRELACIONES entre variables y entre variables y factores

Análisis factorial -*FA*-

El **ANÁLISIS FACTORIAL** (como el análisis de componentes principales), TIENE COMO OBJETIVO REDUCIR LA DIMENSIONALIDAD DE LOS DATOS

El análisis factorial surge del interés por comprender las dimensiones de la inteligencia humana en los años 30 del siglo pasado. Sus orígenes se deben a **Spearman, C.** (1904) Psicólogo inglés. También contribuyeron al mismo de forma significativa **Pearson** y **Hotelling** (1933) **Thurstone**, (1947). Los mayores avances en esta técnica se han producido en el campo de la psicometría.



C. Spearman

El modelo de análisis factorial especifica que las variables vienen determinadas por los FACTORES COMUNES (calculados como en ACP) y por FACTORES ÚNICOS (uno específico para cada variable); las estimaciones calculadas se basan en el supuesto de que ningún factor único está correlacionado con los demás, ni con los factores comunes.

Análisis factorial -*FA*-

El **ANÁLISIS FACTORI**
TIENE COMO OBJETIV

El análisis factorial surge del
la inteligencia humana en los
deben a **Spearman, C.** (19
contribuyeron al mismo de fo
(1933) **Thurstone**, (1947
han producido en el campo d

Las Matemáticas tienen mucho en común con las otras materias de Ciencias, pero tienen algo específico que las diferencia de cualquier otra materia de Ciencias.

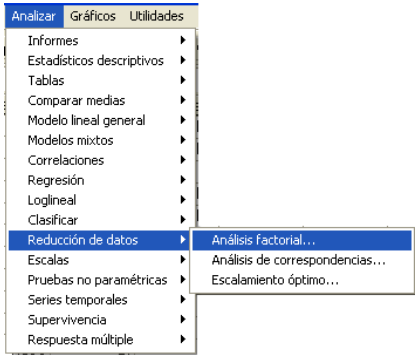
El modelo de Análisis Factorial captura esta información

El modelo de análisis fa especifica que las variables vienen determinadas por los FACTORES COMUNES (calculados como en ACP) y por FACTORES ÚNICOS (uno específico para cada variable); las estimaciones calculadas se basan en el supuesto de que ningún factor único está correlacionado con los demás, ni con los factores comunes.

SPSS

¿Cómo realizar un *AF* en SPSS?

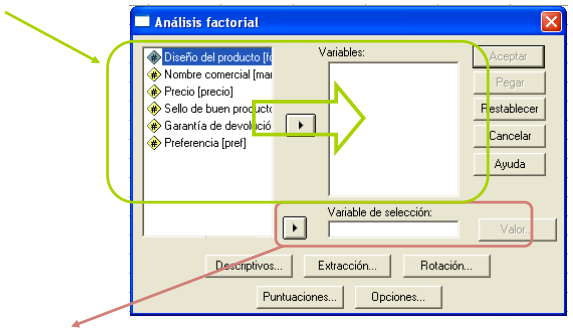
- Para obtener un Análisis Factorial (AF)
  - Seleccione los menús:



SPSS

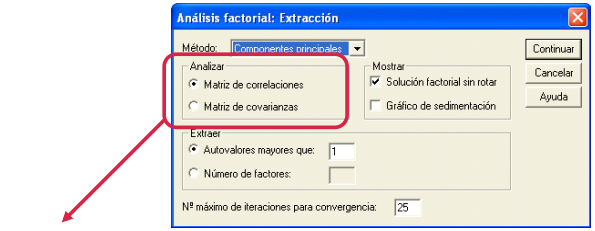
Selección de variables

En primer lugar, se deben seleccionar las variables que intervendrán en el análisis factorial.



Selección de casos en el análisis factorial
Para seleccionar los casos para el análisis, elija una variable de selección y pulse en Valor para introducir un entero como el valor de selección. En el análisis factorial, sólo se usarán los casos con ese valor para la variable de selección.

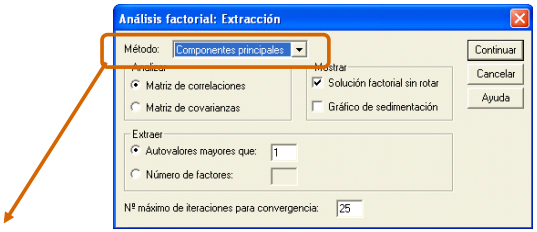
SPSS



Permite especificar o una **MATRIZ DE CORRELACIONES** o de **COVARIANZAS**

SPSS

• EXTRACCIÓN



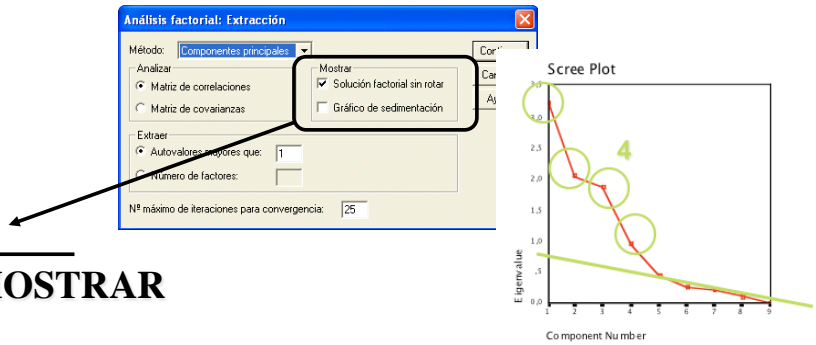
– MÉTODOS

Permite especificar el método de extracción factorial. Los métodos disponibles son:

- **COMPONENTES PRINCIPALES**
- **MÍNIMOS CUADRADOS NO PONDERADOS**
- **MÍNIMOS CUADRADOS GENERALIZADOS**
- **MÁXIMA VEROSIMILITUD**
- **FACTORIZACIÓN DE EJES PRINCIPALES**
- **FACTORIZACIÓN ALFA**
- **FACTORIZACIÓN IMAGEN.**

SPSS

• EXTRACCIÓN



MOSTRAR

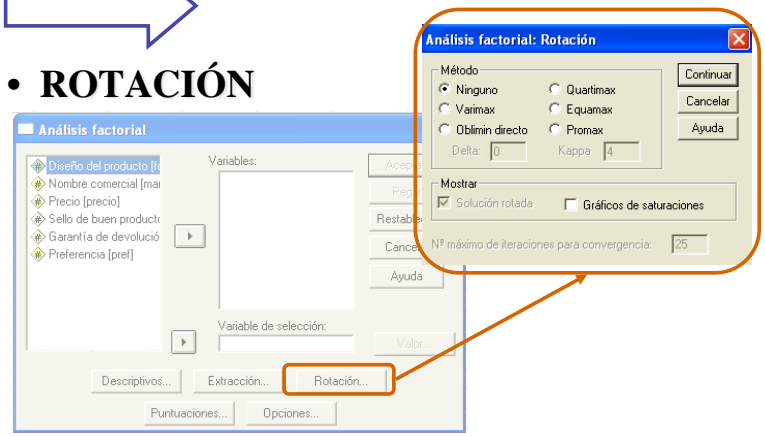
– **Gráfico de sedimentación** de los autovalores (Scree Plot)

Gráfico de la varianza asociada a cada factor. Se utiliza para determinar cuántos factores deben retenerse.

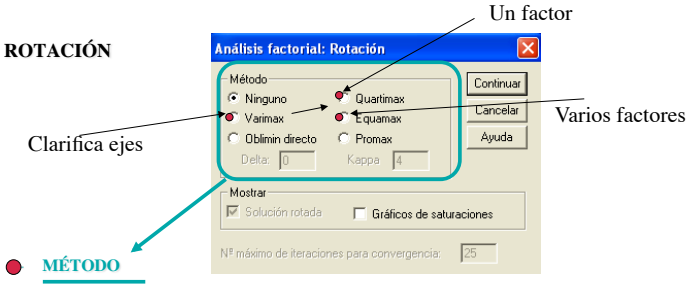
Típicamente el gráfico muestra la clara ruptura entre la pronunciada pendiente de los factores más importantes y el descenso gradual de los restantes (los sedimentos).

SPSS

• ROTACIÓN



- **ROTACIÓN**



MÉTODO

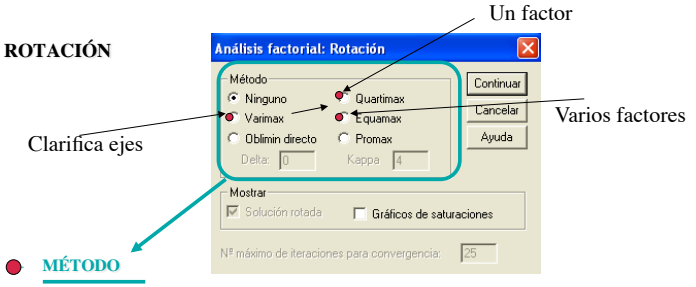
Permite seleccionar el método de rotación factorial. Los métodos disponibles son: **(ortogonales) VARIMAX, QUARTIMAX, EQUAMAX**, y (oblícuos) oblimin directo y promax.

VARIMAX Método de rotación ortogonal que minimiza el número de variables que tienen saturaciones altas en cada factor. Simplifica la interpretación de los factores.

QUARTIMAX Método de rotación que minimiza el número de factores necesarios para explicar cada variable. Simplifica la interpretación de las variables observadas.

EQUAMAX Método de rotación que es combinación del método varima.. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.

- **ROTACIÓN**

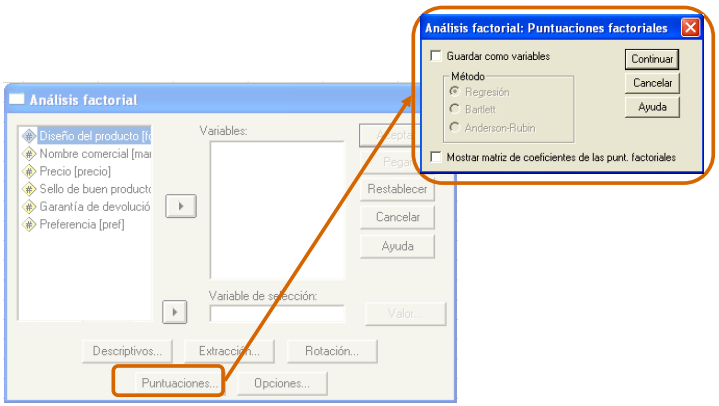


MÉTODO

Equamax: las absorción de varianza se reparte por igual entre los ejes
Quartimax: La mayor parte de la varianza es absorbida por el primer eje
Varimax: Intermedia entre Equamax y Quartimax

variable.

PUNTUACIONES FACTORIALES



Bibliografía

CUADRAS, C.M. (1996). ***Métodos de Análisis Multivariante***, EUB, Barcelona.

HAIR, J.F., ANDERSON,R.E., TATHAM, R.L. and BLACK, W.C. (1998). ***Multivariate Data Analysis***, Prentice Hall,New Jersey.

JOLLIFE, I.T. (1986). ***Principal Component Analysis***, Springer-Verlag, New York.

JOHNSON, D.E. (1998). ***Métodos Multivariados aplicados al análisis de datos***, Thomson Eds., México.

CIC
Centro de Investigación del Cáncer

UNIVERSIDAD DE SALAMANCA
Dpto. de Estadística

BIPLOT methods applied to gene expression data

Dra. Purificación Galindo Villardón
pgalindo@usal.es

ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect
2008
Journal of Statistical Planning and Inference 138 (2008) 500–515

journal of statistical planning and inference
www.elsevier.com/locate/jspi

Several biplot methods applied to gene expression data

Mira Park[a], Jae Won Lee[b,*], Jung Bok Lee[c], Seuck Heun Song[b]

[a]*Department of Pre-medicine, Eulji University, 143-5 Yongdu-dong, Chung-gu, Daejon 301-832, Korea*
[b]*Department of Statistics, Korea University, 5-1 Anam-dong, Seongbuk-gu, Seoul 701-112, Korea*
[c]*Institute of Human Genomic Study, College of Medicine, Korea University, Gojan1-dong, Danwon-gu, Ansan, Gyeonggi-do, Korea*

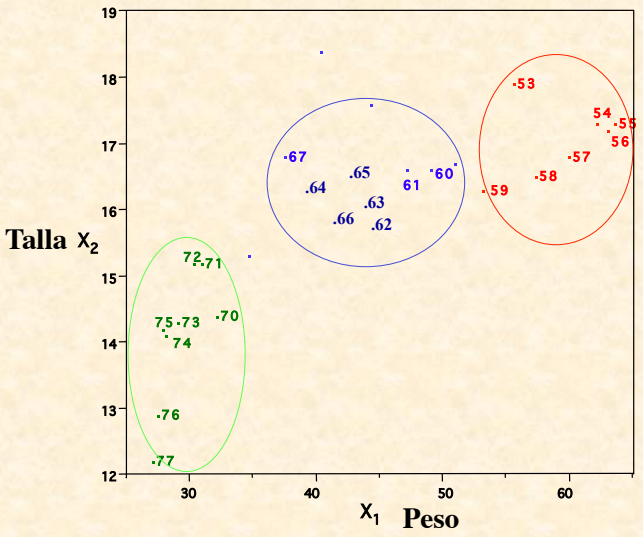
Abstract

DNA microarray experiments result in enormous amount of data, which need careful interpretation. Biplot approaches show simultaneous display of genes and samples in low-dimensional graphs and thus can be used to represent the relationships between genes and samples. There are several different types of biplots, and these methods need to be evaluated because each plot provides different result.

In this paper, we review several variants of biplot methods such as principal component analysis biplot, factor analysis biplot, multidimensional scaling biplot and correspondence analysis biplot. We investigate the properties of these methods and compare their performances by analyzing various types of well-known gene expression data. We also suggest the supplementary data method as a tool for (i) classifying the previously unknown sample/gene to existing class, (ii) analyzing mixture data and (iii) presenting illustrative variables, etc. The usefulness of this approach for interpreting microarray data is demonstrated.

Keywords: Gene expression data; Biplot; Supplementary data; Principal component analysis; Factor analysis; Correspondence analysis; Multidimensional scaling

$\mathbf{X}_{nx2}$



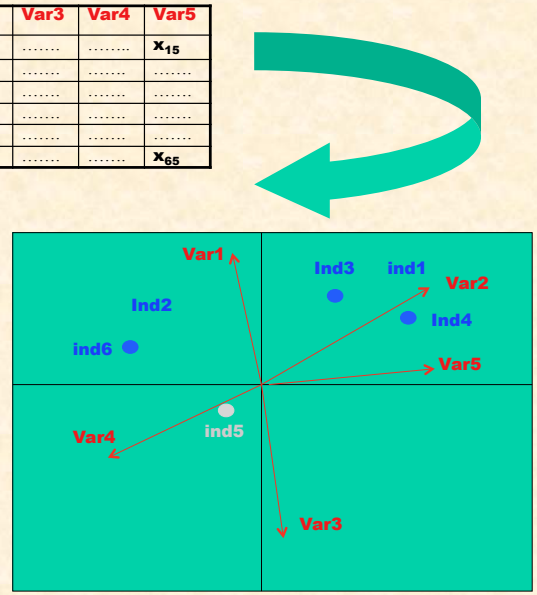
X_{nxp}

- Representación multidimensional: R^p
- Hipernube

Representación en **baja dimensión**

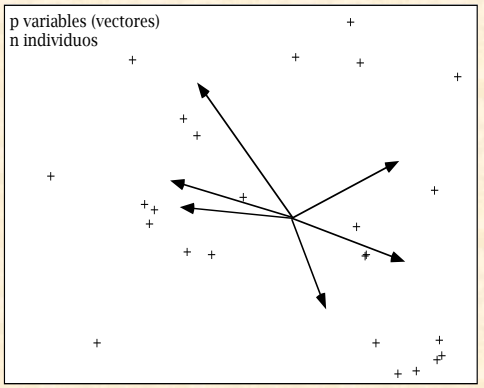
Variables latentes o hipotéticas

	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



BIPLOT: DEFINICIÓN INTUITIVA

Un **BIPLOT** (GABRIEL, 1971) es una representación gráfica de datos multivariantes. De la misma manera que un diagrama de dispersión muestra la distribución conjunta de dos variables, un BIPLOT representa tres o más variables. (GABRIEL y ODOROFF, 1990).



Abstract

DNA microarray experiments result in enormous amount of data, which need careful interpretation. Biplot approaches show simultaneous display of genes and samples in low-dimensional graphs and thus can be used to represent the relationships between genes and samples. There are several different types of biplots, and these methods need to be evaluated because each plot provides different result.

In this paper, we review several variants of biplot methods such as principal component analysis biplot, factor analysis biplot, multidimensional scaling biplot and correspondence analysis biplot. We investigate the properties of these methods and compare their performances by analyzing various types of well-known gene expression data. We also suggest the supplementary data method as a tool for (i) classifying the previously unknown sample/gene to existing class, (ii) analyzing mixture data and (iii) presenting illustrative variables, etc. The usefulness of this approach for interpreting microarray data is demonstrated.

Keywords: Gene expression data; Biplot; Supplementary data; Principal component analysis; Factor analysis; Correspondence analysis; Multidimensional scaling

X

XX^T X^TX

ECKART y YOUNG (1936, 1939)

$X = U D V^T$

U es una matriz cuyos vectores columna son ortonormales y vectores propios de **XX'**.

V es una matriz ortonormal cuyos vectores columna son vectores propios de **X'X**

D es la matriz diagonal de valores singulares de X, que son las raices cuadradas no negativas de los valores propios de **X'X**

AB^T

J=UD K=V	G=U H=VD
JK^T-Biplot	GH^T-Biplot

(Gabriel, 1971)

PCA (1901) — 1900

PCA (1933) (1936)

Biplot Gabriel (1971)

HJ-BIPLOT
(Galindo,1986)

EL HJ-BIPLOT es un representación gráfica multivariante de las líneas de una matriz Xnxp mediante los marcadores $j_1,.....,j_n$ para sus filas y $h_1, ...,h_p$ para sus columnas, elegidos de forma que ambos marcadores puedan ser superpuestos en un mismo sistema de referencia con máxima calidad de representación (Galindo, 1986)

Questiió
Vol 10 Nº1
pp:13-23

UNA ALTERNATIVA DE REPRESENTACIÓN SIMULTÁNEA: HJ-BIPLOT

M.ª PURIFICACIÓN GALINDO VILLARDÓN

UNIVERSIDAD DE SALAMANCA

1986

En el presente trabajo se hace una revisión del método BIPLOT propuesto por GABRIEL y se propone una nueva forma de representación simultánea para matrices de datos que denominamos HJ-BIPLOT, en la cual las coordenadas para las columnas coinciden con los marcadores para las columnas en el GH'-biplot y las coordenadas para las filas coinciden con los marcadores para las filas en el JK'-biplot de GABRIEL. Estas coordenadas pueden ser representadas en un mismo sistema de referencia: El sistema de los ejes factoriales.

Se demuestra que el HJ-biplot consigue la misma bondad de ajuste para filas y para columnas, siendo ésta de un orden muy superior al usual.

Se demuestra también, que el HJ-biplot, para matrices de datos positivos, da lugar a las mejores representaciones Β-baricéntricas.

PCA (1901)

PCA (1933) (1936)

1950

GH/JK Biplot GABRIEL (1971)

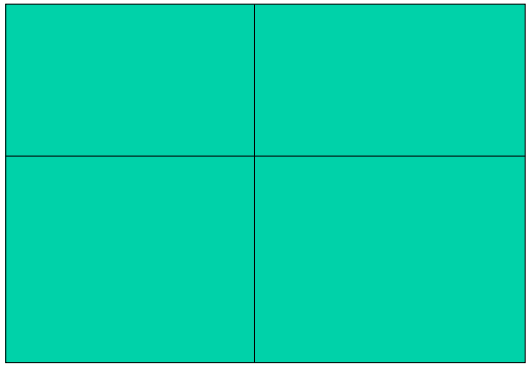
HJ-Biplot Galindo (1986)

2000

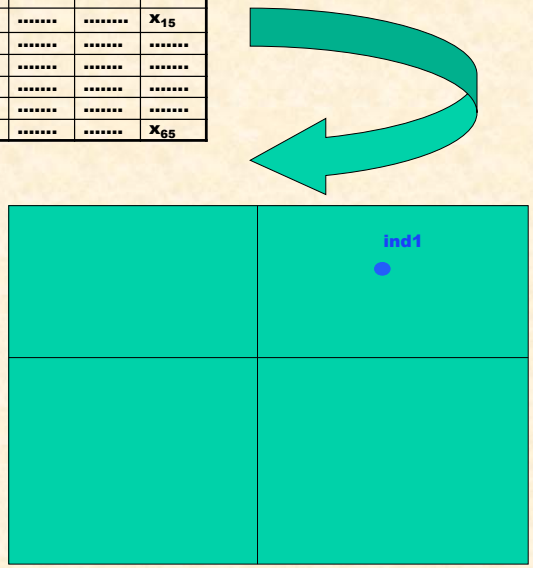
¿Cómo se interpreta un HJ-BIPLOT?

	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}

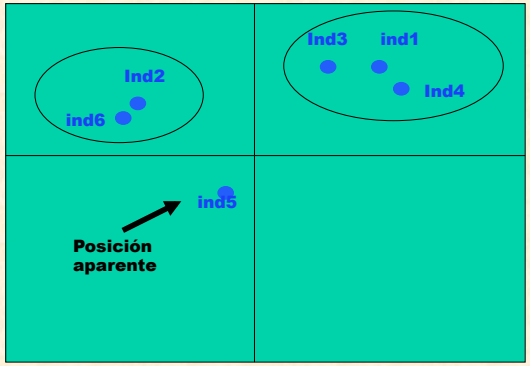
....



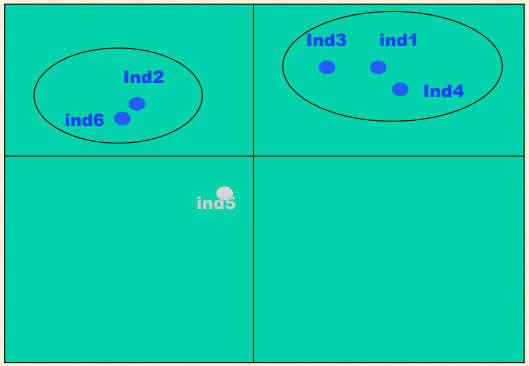
	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



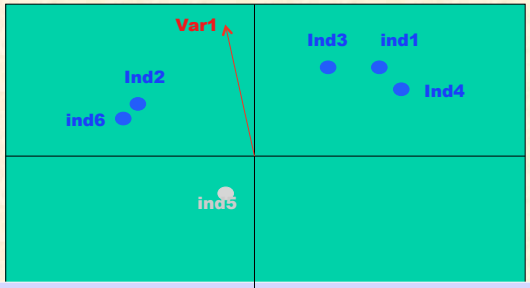
	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



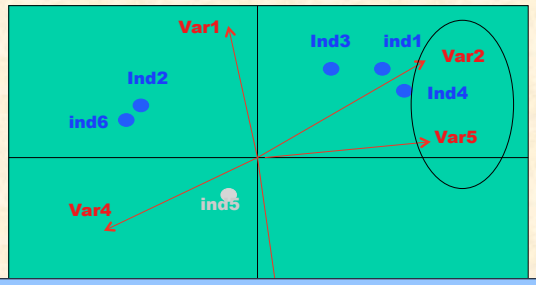
	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



Cuanto más distantes aparezcan los puntos que representan a las variables del centro de gravedad, más variabilidad habrán presentado esos caracteres en el estudio

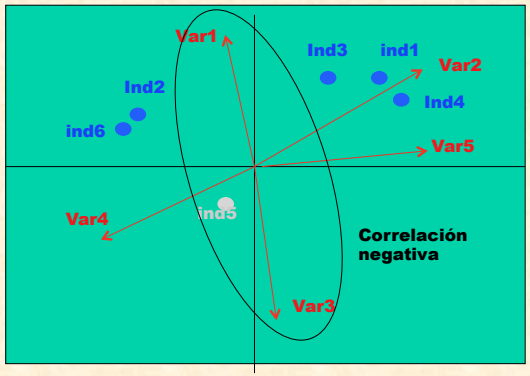
	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}

Correlación Positiva



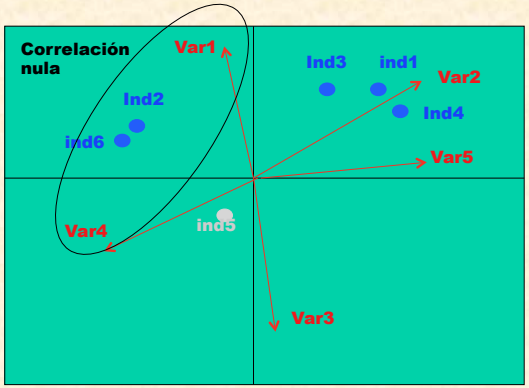
Cuanto menor sea el ángulo que forman dos vectores que unen el centro de gravedad con los puntos que representan a las variables, mayor correlación

	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



ANGULO OBTUSO indica Relación inversa entre las variables

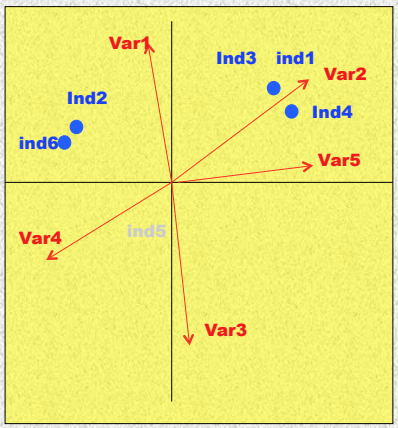
	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



Ángulo recto indica independencia

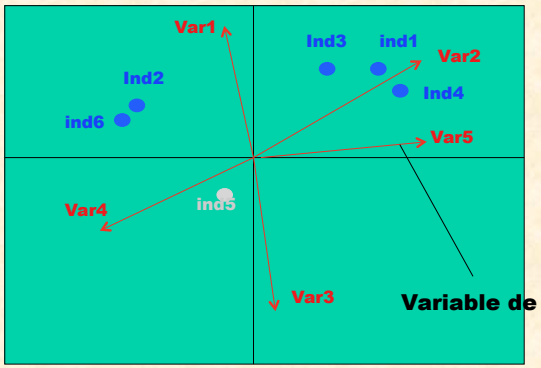
HJ BIPLOT

A partir del gráfico se puede conocer

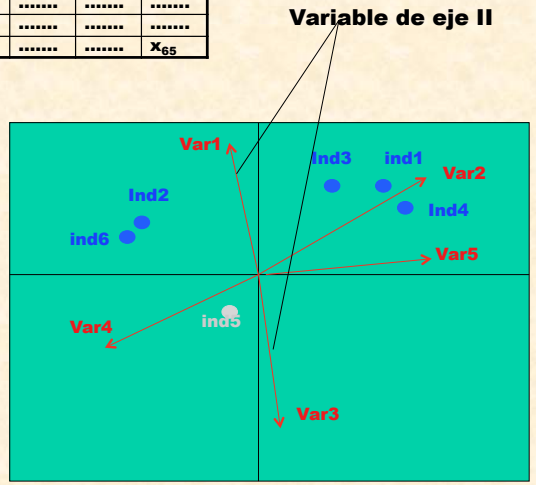


- **La variabilidad de las variables**:
  observando la longitud del vector
- **La covariación de las variables**:
  observando el ángulo
- **La similitud en el patrón de los individuos:**
  analizando su proximidad

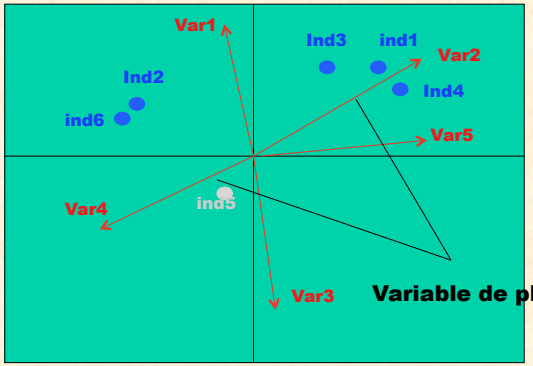
	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
Ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



	Var1	Var2	Var3	Var4	Var5
ind1	x_{11}	.....	.....	.....	x_{15}
ind2	.....	.....	.....	.....	.....
Ind3	.....	.....	.....	.....	.....
Ind4	.....	.....	.....	.....	.....
ind5	.....	.....	.....	.....	.....
ind6	x_{61}	.....	.....	.....	x_{65}



GALINDO, M.P. (1986). 'Una alternativa de representación simultanea: HJ-Biplot'. *Questiio*, 10(1): 13-23.

GALINDO, M.P. & CUADRAS, C.M. (1986). 'Una extensión del método biplot a su relación con otras técnicas'. *Publicación de Bioestadística y Biomatemática*. Universidad de Barcelona. Nº. 17.

SOFTWARE

SOFTWARE

José Luís Vicente Vilardón

Classical Biplot-MultBiplot. Multivariate Analysis using Biplots

http://biplot.dep.usal.es/classicalbiplot/

ALTERNATIVAS DE CÓDIGO ABIERTO

Elisa Frutos Bernal y Mª Purificación Galindo Villardón (2012)

GGEBiplotGUI: Interactive GGE Biplots in R

Para descargar: http://cran.r-project.org/web/packages/GGEBiplotGUI/index.html

¿Cómo citar?: http://cran.r-project.org/web/packages/GGEBiplotGUI/citation.html

Ana Belen Nieto Librero, Nora Bacala y Mª Purificación Galindo Villardón (2011)

multibiplotGUI: Multibiplot Analysis in R

http://cran.r-project.org/web/packages/multibiplotGUI/index.html

Faria, J.C & Demetrio, C. G. B (2011)

BPCA: Biplot of multivariate data based on Principal Compone

Para descargar: http://cran.r-project.org/web/packages/bpca/

¿Cómo citar? http://cran.r-project.org/web/packages/bpca/citatio

R-Forge

The BiplotGUI package homepage

The BiplotGUI package for R makes it easy to construct and interact with biplots.

Biplots

Biplots can be interpreted as graphs in which observations are represented as points while, simultaneously, variables are represented as calibrated biplot axes. Such representations make it easy to visualise multivariate data in two or three dimensions. The biplots of the BiplotGUI package are based on the book by Gower and Hand (1996) and can be thought of as multivariate analogues of the ordinary scatter plot.

http://diarium.usal.es/pgalindo/

¿BIPLOT?

Abstract

DNA microarray experiments result in enormous amount of data, which need careful interpretation. Biplot approaches show simultaneous display of genes and samples in low-dimensional graphs and thus can be used to represent the relationships between genes and samples. There are several different types of biplots, and these methods need to be evaluated because each plot provides different result.

GH-BIPLOT

In this paper, we review CORRESPONDENCIAS al component analysis biplot, factor analysis biplot, multidimensional scaling biplot and correspondence analysis biplot. We investigate the properties of these methods and compare MULTIDIMENSIONAL SCALING ression data. We also suggest the supplementary data method as a tool for (i) classifying the previously unknown sample/gene to existing class, (ii) analyzing mixture data and (iii) presenting illustrative variables, etc. The usefulness of this approach for interpreting microarray data is demonstrated.

PLOT

© 20 sevier B.V. All rights reserv

BI

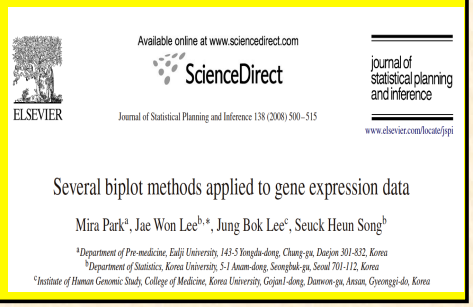
Keywords: Gene expression d inter plem data; Principal component analysis; Factor analysis; Correspondence analysis; Multidimensional scaling

BIBLIOGRAFIA BASE

GALINDO, M.P. (1986). 'Una alternativa de representación simultanea: HJ-Biplot'. *Questiio*, 10(1): 13-23.

GABRIEL, K.R. (1971). 'The biplot-graphic display of matrices with application to principal component analysis'. *Biometrika*, 58: 453-467.

GOWER, J.C. & HAND, D.J. (1996). *Biplots*. Chapman & Hall. London.


Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 500–515
journal of statistical planning and inference
www.elsevier.com/locate/jspi
ELSEVIER
Several biplot methods applied to gene expression data
Mira Park[a], Jae Won Lee[b,*], Jung Bok Lee[c], Seuck Heun Song[b]
[a]Department of Pre-medicine, Eulji University, 143-5 Yongdu-dong, Chung-gu, Daejon 301-832, Korea
[b]Department of Statistics, Korea University, 5-1 Anam-dong, Seongbuk-gu, Seoul 701-112, Korea
[c]Institute of Human Genomic Study, College of Medicine, Korea University, Gojan1-dong, Danwon-gu, Ansan, Gyeonggi-do, Korea

BIPLOT

'Salamanca Statistics Seminar III'
Advances in Descriptive Multivariate Analysis.
1996

Professor Ruben GABRIEL *(University of Rochester, U.S.A.), trying to find a Biplot at the bottom of the well, in the cloister of "Arzobispo Fonseca Palace"(Salamanca University, Spain)*

Purificación GALINDO VILLARDON
pgalindo@usal.es

COLINEALIDAD

Departamento de ESTADÍSTICA
Universidad de Salamanca

ESTIMADORES

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Los coeficientes de regresión :

Los coeficientes de regresión se interpretan como el efecto de la variable X_j en la variable dependiente Y, cuando el resto de las variables se mantienen constantes; es decir, el cambio producido en la variable dependiente Y por cada incremento unitario en la regresora X_j manteniendo constante el resto de las predictoras.

Coeficientes de regresión estandarizados:

Cada $Beta_i$ (estandarizado) se interpreta como el cambio, en unidades de desviación típica, en la variable dependiente, por cada cambio en una desviación típica en la variable independiente Xi, manteniendo el resto de las variables independientes constantes.

Equivale a realizar una regresión sobre datos estándarizados
(media cero, desviación tipica 1)

COLINEALIDAD

- ***Cuando las variables explicativas están relacionadas se dice que hay colinealidad.***
- **En presencia de colinealidad los coeficientes de regresión son inestables y por tanto no son interpretables**

SÍNTOMAS DE COLINEALIDAD

- -1.- **A**ltas correlaciones entre al menos un par de variables.
- -2.- **A**parece como no significativa alguna variable
- que el investigador sabe que es importante.
- -3.- **L**os *errores estándar de los estimadores* son anormalmente grandes, disminuyendo drásticamente al eliminar una o varias variables regresoras.

La colinealidad puede estar presente sin que estos síntomas sean evidentes…

El ususario puede identificar los posibles síntomas de COLINEALIDAD, pero el diagnóstico y el tratamiento requiere la participación de un estadístico experto en Modelos de Regresión

DATOS

Y

CLASE	ALQUITRAN	NICOTINA	PESO	MONÓXIDO
RUBIO	4,10	0,860	0,9853	13,60
NEGRO	16,00	1,060	1,0938	16,60
NEGRO	29,80	2,030	1,1650	23,50
RUBIO	8,00	0,670	0,9280	10,20
RUBIO	4,10	0,400	0,9462	5,40
NEGRO	15,00	1,040	0,8885	15,00
RUBIO	8,80	0,750	1,0267	9,00
RUBIO	12,40	0,950	0,9225	12,30
NEGRO	16,60	1,120	0,9372	16,30
RUBIO	14,90	1,020	0,8858	15,40
RUBIO	13,70	1,010	0,9643	13,00
RUBIO	15,10	0,900	0,9316	14,40
RUBIO	7,80	0,570	0,9705	10,00
RUBIO	11,40	0,780	1,1240	10,20
RUBIO	9,00	0,740	0,8517	9,50
RUBIO	1,00	0,130	0,7851	1,50
NEGRO	17,00	1,260	0,9186	18,50
RUBIO	12,80	1,080	1,0395	12,60
NEGRO	15,80	0,960	0,9573	17,50
RUBIO	4,50	0,420	0,9106	4,90
NEGRO	14,50	1,010	1,0070	15,90
RUBIO	7,30	0,610	0,9806	8,50
NEGRO	8,60	0,690	0,9693	10,60
NEGRO	15,20	1,020	0,9496	13,90
RUBIO	12,00	0,820	1,1184	14,90

The United States Surgeon General, considera cada una de estas sustancias peligrosas para la salud de los fumadores. Estudios ya realizados, ponen de manifiesto que **incrementos en el contenido de alquitrán y nicotina de los cigarrillos vienen acompañados por incrementos en el monóxido de carbono** emitido al fumar.

La Variable dependiente es **Y: *contenido en monóxido de carbono***

Las regresoras (variables explicativas) son:

X_1 : ***contenido de alquitrán,***
X_2 : ***contenido en nicotina,***
X_3: ***peso del cigarrillo***
X_4: ***clase de tabaco***

OBJETIVO:
Encontrar un modelo que nos permita estimar la cantidad de MONOXIDO de CARBONO, a partir de las variables Alquitrán, Nicotina y Peso del cigarrillo

"Using Cigarette Data for an Introduction to Multiple Regression", by Lauren Mcintyre in Volume 2, Number 1, of the *Journal of Statistics Education*.

OBJETIVO:
Encontrar un modelo que nos permita estimar la cantidad de MONOXIDO de CARBONO, a partir de las variables Alquitrán, Nicotina y Peso del cigarrillo

MODELOS UNIVARIANTES

Modelo M_0 — *Monox. Carbono* = 12.53

Modelo M_A — *Monox. Carb= 3.88+0.73 Alquitrán* ***R^2= 0.82***

Modelo M_N — *Monox. Carb= 1.67+12.40 Nicotina* ***R^2= 0.86***

Modelo M_P — *Monox. Carb= -11.80+ 25.06 Peso* ***R^2= 0.215***

Modelo M_{ANP} — MODELO MULTIVARIANTE

$$Carbono = \beta_0 + \beta_1 Alquitrán + \beta_2 Nicotina + \beta_3 Peso$$

Y

CLASE	ALQUITRAN	NICOTINA	PESO	MONÓXIDO
RUBIO	4,10	0,860	0,9853	13,60
NEGRO	16,00	1,060	1,0938	16,60
NEGRO	29,80	2,030	1,1650	23,50
RUBIO	8,00	0,670	0,9280	10,20
RUBIO	4,10	0,400	0,9462	5,40
NEGRO	15,00	1,040	0,8885	15,00
RUBIO	8,80	0,750	1,0267	9,00
RUBIO	12,40	0,950	0,9225	12,30
NEGRO	16,60	1,120	0,9372	16,30
RUBIO	14,90	1,020	0,8858	15,40
RUBIO	13,70	1,010	0,9643	13,00
RUBIO	15,10	0,900	0,9316	14,40
RUBIO	7,80	0,570	0,9705	10,00
RUBIO	11,40	0,780	1,1240	10,20
RUBIO	9,00	0,740	0,8517	9,50
RUBIO	1,00	0,130	0,7851	1,50
NEGRO	17,00	1,260	0,9186	18,50
RUBIO	12,80	1,080	1,0395	12,60
NEGRO	15,80	0,960	0,9573	17,50
RUBIO	4,50	0,420	0,9106	4,90
NEGRO	14,50	1,010	1,0070	15,90
RUBIO	7,30	0,610	0,9806	8,50
NEGRO	8,60	0,690	0,9693	10,60
NEGRO	15,20	1,020	0,9496	13,90
RUBIO	12,00	0,820	1,1184	14,90

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1,838	4,388		,419	,680
	alquitrán	,239	,196	,296	1,217	,237
	nicotina	8,612	3,332	,644	2,584	,017
	peso	,334	4,954	,006	,067	,947

a. Dependent Variable: monox.carb.

Y

CLASE	ALQUITRAN	NICOTINA	PESO	MONÓXIDO
RUBIO	4,10	0,860	0,9853	13,60
NEGRO	16,00	1,060	1,0938	16,60
NEGRO	29,80	2,030	1,1650	23,50
RUBIO	8,00	0,670	0,9280	10,20
RUBIO	4,10	0,400	0,9462	5,40
NEGRO	15,00	1,040	0,8885	15,00
RUBIO	8,80	0,750	1,0267	9,00
RUBIO	12,40	0,950	0,9225	12,30
NEGRO	16,60	1,120	0,9372	16,30
RUBIO	14,90	1,020	0,8858	15,40
RUBIO	13,70	1,010	0,9643	13,00
RUBIO	15,10	0,900	0,9316	14,40
RUBIO	7,80	0,570	0,9705	10,00
RUBIO	11,40	0,780	1,1240	10,20
RUBIO	9,00	0,740	0,8517	9,50
RUBIO	1,00	0,130	0,7851	1,50
NEGRO	17,00	1,260	0,9186	18,50
RUBIO	12,80	1,080	1,0395	12,60
NEGRO	15,80	0,960	0,9573	17,50
RUBIO	4,50	0,420	0,9106	4,90
NEGRO	14,50	1,010	1,0070	15,90
RUBIO	7,30	0,610	0,9806	8,50
NEGRO	8,60	0,690	0,9693	10,60
NEGRO	15,20	1,020	0,9496	13,90
RUBIO	12,00	0,820	1,1184	14,90

¿COLINEALIDAD?

Correlations

		alquitrán	nicotina	peso
alquitrán	Pearson Correlation	1,000	,945	,461
	Sig. (2-tailed)		,000	,020
	N	25,000	25	25
nicotina	Pearson Correlation	,945	1,000	,499
	Sig. (2-tailed)	,000		,011
	N	25	25	25
peso	Pearson Correlation	,461	,499	1,000
	Sig. (2-tailed)	,020	,011	
	N	25	25	25,000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Diagnostico Colinealidad:

	Analizar
¿Hay colinealidad?	Condition Number
¿Hasta donde alcanza?	Condition Index
¿Qué coeficientes están afectados?	VIF
¿Qué regresoras están involucradas?	Contribuciones de cada componente al factor de inflación de varianza

DATOS

Y

CLASE	ALQUITRAN	NICOTINA	PESO	MONÓXIDO
RUBIO	4,10	0,860	0,9853	13,60
NEGRO	16,00	1,060	1,0938	16,60
NEGRO	29,80	2,030	1,1650	23,50
RUBIO	8,00	0,670	0,9280	10,20
RUBIO	4,10	0,400	0,9462	5,40
NEGRO	15,00	1,040	0,8885	15,00
RUBIO	8,80	0,750	1,0267	9,00
RUBIO	12,40	0,950	0,9225	12,30
NEGRO	16,60	1,120	0,9372	16,30
RUBIO	14,90	1,020	0,8858	15,40
RUBIO	13,70	1,010	0,9643	13,00
RUBIO	15,10	0,900	0,9316	14,40
RUBIO	7,80	0,570	0,9705	10,00
RUBIO	11,40	0,780	1,1240	10,20
RUBIO	9,00	0,740	0,8517	9,50
RUBIO	1,00	0,130	0,7851	1,50
NEGRO	17,00	1,260	0,9186	18,50
RUBIO	12,80	1,080	1,0395	12,60
RUBIO	15,80	0,960	0,9573	17,50
RUBIO	4,50	0,420	0,9106	4,90
NEGRO	14,50	1,010	1,0070	15,90
RUBIO	7,30	0,610	0,9806	8,50
NEGRO	8,60	0,690	0,9693	10,60
NEGRO	15,20	1,020	0,9496	13,90
RUBIO	12,00	0,820	1,1184	14,90

SPSS

Analyze Graphs Utilities Add-ons Window Help

Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks

Linear...
Curve Estimation...
Partial Least Squares...

Dependent: monox.carb. [MONÓXIDO]
Block 1 of 1
Previous
Independent(s): alquitrán [ALQUITRAN], nicotina [NICOTINA], peso [PESO]

Regression Coefficients: Estimates, Confidence intervals, Covariance matrix
Model fit, R squared change, Descriptives, Part and partial correlations, Collinearity diagnostics

¿Hay colinealidad?

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	alquitrán	nicotina	peso
1	1	3,834	1,000	,00	,00	,00	,00
	2	,154	4,993	,01	,05	,01	,01
	3	,009	21,190	,02	,95	,94	,00
	4	,003	34,509	,97	,00	,04	,99

a. Dependent Variable: monox.carb.

Cond Number>30

Regresoras involucradas

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	1,838	4,388		,419	,680	-7,287	10,964		
	alquitrán	,239	,196	,296	1,217	,237	-,169	,647	,107	9,382
	nicotina	8,612	3,332	,644	2,584	,017	1,682	15,542	,102	9,841
	peso	,334	4,954	,006	,067	,947	-9,968	10,637	,750	1,334

a. Dependent Variable: monox.carb.

$$FIV_i = \frac{1}{1 - R_i^2}$$

$$T_i = \frac{1}{FIV_i} = 1 - R_i^2$$

¿Afectadas? VIF>10

SPSS

¿Hay colinealidad?

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	alquitrán	nicotina	peso
1	1	3,834	1,000	,00	,00	,00	,00
	2	,154	4,993	,01	,05	,01	,01
	3	,009	21,190	,02	,95	,94	,00
	4	,003	34,509	,97	,00	,04	,99

a. Dependent Variable: monox.carb.

SACAMOS EL ALQUITRAN

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	nicotina	peso
1	1	2,912	1,000	,00	,01	,00
	2	,084	5,874	,02	,81	,01
	3	,003	30,049	,98	,18	,99

a. Dependent Variable: monox.carb.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	1,838	4,388		,419	,680	-7,287	10,964		
	alquitrán	,239	,196	,296	1,217	,237	-,169	,647	,107	9,382
	nicotina	8,612	3,332	,644	2,584	,017	1,682	15,542	,102	9,841
	peso	,334	4,954	,006	,067	,947	-9,968	10,637	,750	1,334

a. Dependent Variable: monox.carb.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	1,578	4,431		,356	,725	-7,611	10,767		
	nicotina	12,384	1,239	,925	9,992	,000	9,814	14,955	,751	1,332
	peso	,105	5,004	,002	,021	,984	-10,274	10,483	,751	1,332

a. Dependent Variable: monox.carb.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,931[a]	,868	,849	1,84332

a. Predictors: (Constant), peso, alquitrán, nicotina

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,926[a]	,858	,845	1,86341

a. Predictors: (Constant), peso, nicotina

Tratamiento de la colinealidad

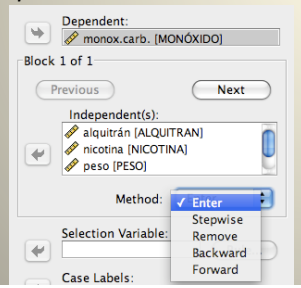
- MÉTODOS PASO A PASO ?

Forward Selection (FS)

Backward Elimination (BE)

Stepwise

- Regresión sobre componentes principales
- OTRAS alternativas



The default tolerance level is 0.0001

Forward Selection

2 *Las variables Nicotina, Alquitrán y Peso explican el 85.2% de la información*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	F Change	df1	df2	Sig. F Change
1	,926	,858	,852	1,82247	,858	139,326	1	23	,000

a. Predictors: (Constant), nicotina

ANOVA

P-valor <0.05

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	462,758	1	462,758	139,326	,000
	Residual	76,392	23	3,321		
	Total	539,150	24			

1 *Los datos contienen información*

a. Predictors: (Constant), nicotina
b. Dependent Variable: monox.carb.

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	1,668	,990		1,686	,105	-,379	3,715
	nicotina	12,397	1,050	,926	11,804	,000	10,224	14,570

a. Dependent Variable: monox.carb.

4 *El modelo es:*

Monx Carb=1.67+12.40 Nicotina

Excluded Variables

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	alquitrán	,295	1,244	,227	,256	,107
	peso	,002	,021	,984	,004	,751

a. Predictors in the Model: (Constant), nicotina
b. Dependent Variable: monox.carb.

3 *El alquitrán y el peso "no explican" nada diferente de lo que ya explica la **nicotina***

Meta-Análisis

Purificación GALINDO VILLARDÓN
pgalindo@usal.es

CIC Centro de Investigación del Cáncer

Universidad de Salamanca

META-ANÁLISIS

Análisis estadístico de los resultados de estudios individuales con el fin de integrar los resultados

- ***Combina información estadística de estudios similares***
- ***Realiza análisis estadísticos de los resultados generales***

Medicina Basada en la Evidencia

META-ANÁLISIS

Utilización de la mejor evidencia científica disponible para tomar decisiones sobre el cuidado de pacientes individuales integrando la experiencia clínica individual

Medicina Basada en la Evidencia

Jano 1997. LIII (1218): 71-72

Niveles de evidencia

Escala de Sackett — De **1** (mejor evidencia) a **5**

1

- RS con homogeneidad de EAC:Ensayos clínicos controlados y randomizados de alta calidad (doble ciego, analizados según intención de tartar…)
- RS con heterogeneidad de EAC de alta calidad
- RS con homogeneidad de EAC de menor calidad
- EAC individuales
- Cohortes
- Casos y Controles
- Serie de casos clínicos
- Opinión de expertos

¿Qué se entiende por Meta-Análisis?

El fundamento de la ciencia se basa en la acumulación de conocimientos.

INVESTIGACIÓN PASADA ⇒ **INVESTIGACIÓN FUTURA**

REVISIONES

TRADICIONALES.

- Subjetividad
- Dificultad en englobar resultados cuantitativos
- Falta de herramientas para sustraer grandes cantidades de información.

META-ANÁLISIS (GLASS, 1976)

"El conjunto de métodos cuantitativos para la comparación y combinación de resultados a partir de estudios analíticos y similares".

Pooling (Peto y cols. 1977)

Overview (Yusuf y cols., 1985)

Análisis cuantitativo de domínios de investigación (Rosenthal, 1984)

Revisión Cuantitativa (Green & Hall, 1984)

ETAPAS DE UN META-ANÁLISIS

ROSENTHAL ,1984

En los estudios meta-analíticos se establecen una serie de etapas las cuales deben ser minuciosamente controladas para desarrollar correctamente la revisión cuantitativa.

1.- Formulación del problema.

2.- Búsqueda de la literatura.

3.- Codificación de los estudios.

4.- Medida de los resultados.

5.- Análisis e interpretación de resultados.

FORMULACIÓN DEL PROBLEMA

✓ **SE DEFINE UN OBJETIVO PRIMARIO EN EL CUAL SE HACE REFERENCIA AL IMPACTO QUE PRODUCE CIERTA INTERVENCION EN UN CONTEXTO DADO.**

✓**SE PRESENTAN UN CONJUNTO DE OBJETIVOS SECUNDARIOS RELACIONADOS QUE AYUDAN A LA COMPRENSION DEL PROBLEMA, INTERACCIONES, ESTUDIOS DE SUBPOBLACIONES**

PROTOCOLO

- **Introduccion del tema**
- **Definicion de los objetivos**
- **Definición de los criterios de inclusion de estudios.**
- **Identificacion de las fuentes de busqueda**
- **Descripción de la información extraible.**
- **Descripción de los análisis a realizar.**

Un ejemplo… FORMULACIÓN DEL PROBLEMA

Objetivo primario

Investigar la eficacia de sedación en términos de calidad, seguridad y coste del tratamiento de midazolam en comparación con propofol a partir de los ensayos aleatorizados que comparan ambos fármacos en pacientes con ventilación mecánica.

Objetivo Secundario

Además identificar posibles subgrupos de pacientes, así como posibles interacciones definidos por las siguientes factores:

- Duración de la sedación (Corta: <24h; Larga:>24h.)
- Tipo de afección del paciente (quirurgica, medica; mixta.

Criterios de inclusión

L os pacientes deben ser asignados a los dos grupos experimentales (propofol/midazolam) de forma aleatoria.

Los ensayos deben valorar alguna de las variables respuesta.

Los pacientes deben estar ventilados mecánicamente en UCI.

Variables medidas

- Tiempo de recuperación después de la sedación:
  - Tiempo en despertar
  - Tiempo hasta la desconexión
  - Tiempo de estancia en UCI.
- Porcentaje del tiempo de sedación en el nivel deseado.
- Valoración de la calidad de sedación
- Ajustes en las dosis para la obtención de la sedación ideal. (Tanto de los respectivos sedantes como otros fármacos administrados conjuntamente)
- Variables hemodinámicas (Presión arterial (Sistólica y diastólica), índice cardiaco, frecuencia cardiaca)
- Medidas respiratorias ($PaCO_2$, PaO_2, VO_2, VCO_2).
- Indices analíticos:
  - función hepática: Transaminasas (GOT, GPT), Bilirubina y Lípidos.
  - Colesterol
  - Triglicérido
- Número de fracasos terapéuticos:debidos tanto a la ineficacia de sedación como a sus complicaciones (Hipotension, incremento en el nivel de Triglicéridos).
- Mortalidad
- Coste
  - Coste de adquisición de los fármacos
  - Coste de cuidados del paciente con los dos fármacos.
  - Coste total de los fármacos.

ETAPAS DE UN META-ANÁLISIS

En los estudios meta-analíticos se establecen una serie de etapas las cuales deben ser minuciosamente controladas para desarrollar correctamente la revisión cuantitativa.

1.- Formulación del problema.

2.- Búsqueda de la literatura.

3.- Codificación de los estudios.

4.- Medida de los resultados.

5.- Análisis e interpretación de resultados.

BUSQUEDA DE LA INFORMACIÓN

Servicios de Abstracts ➡

Permite localizar estudios asociados a determinadas palabras clave en las bases de datos correspondientes al campo de la revisión: Medline; Sociofile; Biological Abstract; Psychofile;etc

Fuentes primarias ➡

Procedimiento Descendente

Consiste en localizar investigaciones previas que figuran en listas de referencias de artículos disponibles por el revisor.

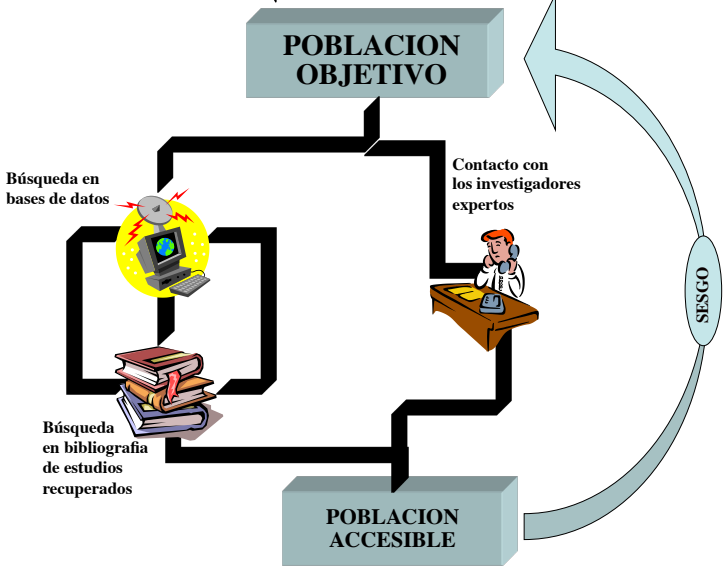
Procedimiento Ascendente

Consiste en localizar investigaciones las cuales citan los trabajos recuperados: índices de citas.

Fuentes Informales ➡

El revisor debe intentar contactar con los investigadores expertos en el tema revisado para la obtención del material bibliográfico que pueda proporcionarle. Es un método interesante para la obtención de artículos no publicados.

BUSQUEDA DE LA INFORMACIÓN



SESGO DE PUBLICACIÓN

EVIDENCIA CLARA DE QUE LOS ESTUDIOS SIGNIFICATIVOS SON MÁS PROBABLES DE PUBLICAR QUE LOS NO SIGNIFICATIVOS.

WAGNER (1986) DEMOSTRÓ QUE TANTO LA DECISIÓN DE ENVIAR, COMO LA VALORACIÓN DE LOS ESTUDIOS, ESTABA ASOCIADA CON LOS RESULTADOS DE LA PUBLICACION.

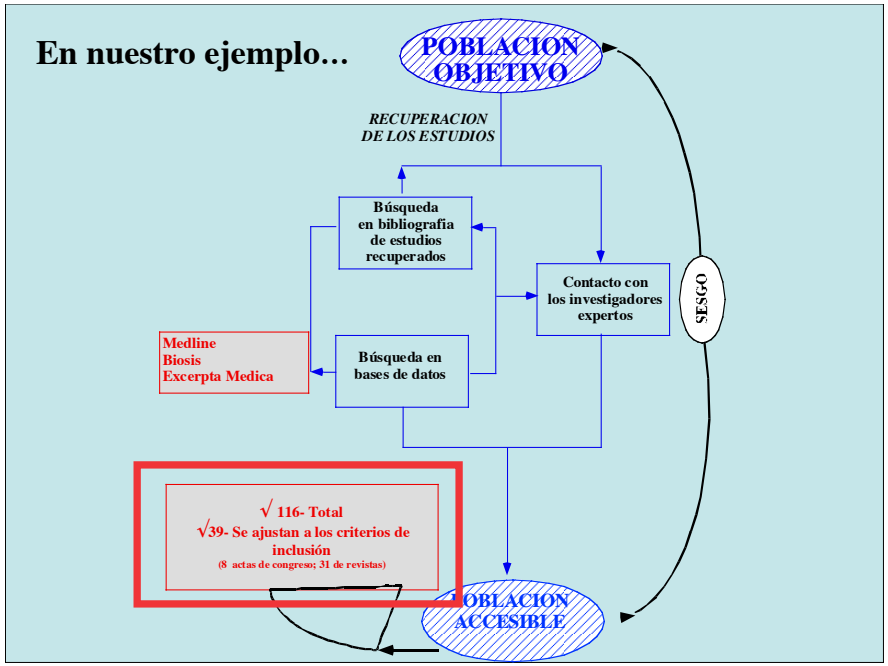
ENVIADOS PARA ACEPTACIÓN
DE LOS ESTUDIOS SIGNIFICATIVOS: 82%.
DE LOS ESTUDIOS NO SIGNIFICATIVOS. 43%.

LOS ARTICULOS ADMITIDOS PARA REVISAR.
DE LOS SIGNIFICATIVOS: 80% FUERON ACEPTADOS.
DE LOS NO SIGNIFICATIVOS: 50% FUERON ACEPTADOS.

Recogida de la información relevante que **aporte evidencia** en relación al objetivo

No suele haber más de 30-40 artículos

En nuestro ejemplo…



ETAPAS DE UN META-ANÁLISIS

En los estudios meta-analíticos se establecen una serie de etapas las cuales deben ser minuciosamente controladas para desarrollar correctamente la revisión cuantitativa.

1.- Formulación del problema.

2.- Búsqueda de la literatura.

3.- Codificación de los estudios.

4.- Medida de los resultados.

5.- Análisis e interpretación de resultados.

ETAPA DE CODIFICACION

OBJETIVOS

√ ***Recoger y estudiar la información de los estudios que puede estar relacionada con los resultados cuantitativos que pueden explicar los resultados finales***

√ ***Valorar los estudios en términos de "calidad"***

ETAPA DE CODIFICACION

PROCESO DE CODIFICACIÓN

-Elaboración de un cuestionario

- Elección de jueces

Características Objetivas

- Número de pacientes en cada estudio
- Año de publicación del estudio
- Criterios de inclusión/exclusión
- Tipo de cegado
- Número de hombres/mujeres
- Edad de los pacientes (Media, Desviación típica, …)
- etc

Características Subjetivas

- Valoración de los análisis estadísticos
- Valoración del diseño
- Valoración de la presentación de resultados
- Adecuación del tipo de cegado
- etc

CODIFICACIÓN

⇒Características Objetivas

- Número de pacientes en cada grupo
- Año de publicación del estudio
- Número de mujeres/hombres
- Edad en años (media/desviación típica o rango)
- Peso medio de los pacientes
- Existencia de criterios de exclusión de pacientes (Si/No)
- Tipo de cegado del diseño (Doble ciego/ Simple ciego/ No ciego)
- Proporción de pacientes excluidos del ensayo.
- Dosis de los diferentes fármacos
- Utilización de otro tipo de sedación utilizada conjuntamente con los sedantes de interés (Anagélsicos/Morfínicos/Otro tipo de sedación)

⇒ Características Subjetivas

- Valoración de los análisis estadísticos
- Valoración del diseño del estudio.
- Valoración de la presentación de los datos.

En nuestro ejemplo…

ETAPA DE CODIFICACION

- **Diferentes valoraciones de los jueces**
- **Errónea definición de los ítems.**

Protección de la fiabilidad

Definición de las características

Normas y reglas de codificación

Aplicación de pruebas piloto para la detección de cuestiones mal planteadas o ambiguas

Entrenamiento de los codificadores para la perfecta interpretación de los ítems.

Análisis de la fiabilidad

Spearman-Brown (α-Cronbach)

$$R = \frac{nr}{1 + (n-1)r}$$

n= nº de codificadores
r= es la media entre los n.(n-1)/2 coeficientes de correlación

$$R = \frac{MC_{entre\,estudios} - MC_{error}}{MC_{entre\,estudios}}$$

Coeficiente Kappa

ETAPAS DE UN META-ANÁLISIS

En los estudios meta-analíticos se establecen una serie de etapas las cuales deben ser minuciosamente controladas para desarrollar correctamente la revisión cuantitativa.

1.- Formulación del problema.

2.- Búsqueda de la literatura.

3.- Codificación de los estudios.

4.- Medida de los resultados.

5.- Análisis e interpretación de resultados.

MEDIDA DE LOS RESULTADOS

Para poder integrar los resultados de un conjunto de estudios es preciso que se expresen en una escala de medida común. En general, las variables dependientes empleadas en los estudios difieren entre sí, razón por la cual, no son directamente comparables.

Permite determinar si el efecto es distinto de cero

P-VALOR

Permite determinar "en qué medida" el efecto es distinto de cero

TAMAÑO DEL EFECTO

Tamaño del Efecto

Variable Respuesta

Variable explicativa	Cuantitativa	Dicotómica	Tiempo
Cuantitativa	❑Correlación		❑Razón de Riesgos
Dicotómica	❑Diferencia de medias ❑Diferencia de medias Estandarizada	❑Riesgo Relativo ❑Riesgo Absoluto ❑Odds-Ratio	

M. Independientes

Estimación tamaño del efecto

Diferencia de medias

$$\delta = \mu_T - \mu_C$$

Estimador $\hat{\delta} = \bar{x}_T - \bar{x}_C$

Varianza $\sigma^2_{\hat{\delta}} = \frac{\sigma^2_T}{n_T} + \frac{\sigma^2_C}{n_C}$

M. Independientes

Estimación tamaño del efecto

Diferencia de medias estandarizada

Cohen 1967

$$\delta = \frac{\mu_T - \mu_C}{\sigma}$$

Estimadores

GLASS 1976

$$d_B = \frac{\bar{x}_T - \bar{x}_C}{S_C}$$

Sesgado

HEDGES 1981

$$d_B = \frac{\bar{x}_T - \bar{x}_C}{S^*}$$

$$s^* = \sqrt{\frac{(n_T-1)s_T^2 + (n_C-1)s_C^2}{n_T + n_C - 2}}$$

Sesgado

HEDGES 1981

$$d_U = c(m)d_B$$

$$c(m) = \frac{\Gamma(\frac{m}{2})}{\sqrt{\frac{m}{2}}\Gamma(\frac{m-1}{2})} \approx 1 - \frac{3}{4m-1}$$

$$m = n_T + n_C - 2$$

Insesgado

Estudios de simulación (HEDGES, 1982) para conocer el funcionamiento de los estimadores de Hedges, han puesto de manifiesto que:

El estimador insesgado d_U proporciona mejores estimaciones cuando el tamaño del efecto es grande, o cuando el número de individuos en cada estudio es pequeño.

Cuando el tamaño del efecto es moderado o pequeño y/o los tamaños muestrales son grandes, el sesgo de d_B es despreciable

Tamaño del Efecto

Variable Respuesta

Variable explicativa	Cuantitativa	Dicotómica	Tiempo
Cuantitativa	✓Correlación		❑Razón de Riesgos
Dicotómica	✓Diferencia de medias ✓Diferencia de medias Estandarizada	❑Riesgo Relativo ❑Riesgo Absoluto ❑Odds-Ratio	

Tablas 2x2

	E (+)	E (-)
FR (+)	**a**	**b**
FR (-)	**c**	**d**

$$P(E+/FR+) = \frac{a}{a+b} = p_{E+/FR+}$$

$$P(E+/FR-) = \frac{c}{c+d} = p_{E+/FR-}$$

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{p_{E+/FR+}}{p_{E+/FR-}}$$

$$P(E+/FR+) = \frac{a}{a+b}$$

$$P(E-/FR+) = \frac{b}{a+b}$$

$$O_1 = \frac{a}{b} = \frac{p_{E+/FR+}}{p_{E-/FR+}}$$

$$P(E+/FR-) = \frac{c}{c+d}$$

$$P(E-/FR-) = \frac{d}{c+d}$$

$$O_2 = \frac{c}{d} = \frac{p_{E+/FR-}}{p_{E-/FR-}}$$

$$OR = \frac{O_1}{O_2} = \frac{a.d}{b.c}$$

$$P(E+/FR+) = \frac{a}{a+b} = p_{E+/FR+}$$

$$P(E+/FR-) = \frac{c}{c+d} = p_{E+/FR-}$$

$$p_+ - p_-$$

TE | **VARIANZA**

$$\ln RR \quad Var(\ln RR) = \frac{b}{(a+b).a} + \frac{d}{(c+d).c}$$

$$\ln OR \quad Var(\ln OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

$$\hat{\Delta} = p_+ - p_- \quad \hat{\sigma}^2_{\hat{\Delta}} = \frac{p_+(1-p_+)}{n_{FR+}} + \frac{p_-(1-p_-)}{n_{FR-}}$$

En nuestro ejemplo..

MEDIDA DE LOS RESULTADOS

Tamaño del Efecto

- **Odds-ratio.**

- **Diferencia de Medias Estandarizada.**

$$\delta = \frac{\mu_T - \mu_C}{\sigma}$$

	Estimador	Varianza
Sesgado	$d_B = \frac{\bar{X}_T - \bar{X}_C}{S^*}$	$\sigma^2_{d_B} = \frac{n_T + n_C}{n_T n_C}$
Insesgado	$d_U = c(m)g$ $c(m) = \frac{\Gamma(\frac{m}{2})}{\sqrt{\frac{m}{2}}\Gamma(\frac{m-1}{2})} \approx 1 - \frac{3}{4m-1}$	$\sigma^2_{d_U} = \frac{n_T + n_C}{n_T n_C} + \frac{\delta^2}{2(n_T + n_C)}$

ETAPAS DE UN META-ANÁLISIS

En los estudios meta-analíticos se establecen una serie de etapas las cuales deben ser minuciosamente controladas para desarrollar correctamente la revisión cuantitativa.

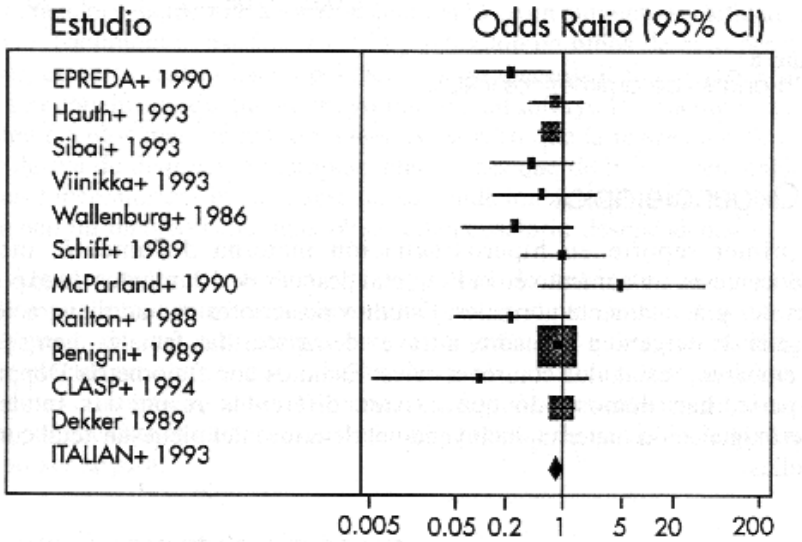
1.- Formulación del problema.

2.- Búsqueda de la literatura.

3.- Codificación de los estudios.

4.- Medida de los resultados.

5.- Análisis e interpretación de resultados.



Autores (p.o. de firma): **M.P. GALINDO**.
Título: Meta-Análisis: El Uso de los Métodos Estadísticos en Revisiones de Estudios de Investigación Relacionados.

Ref. ☐ revista: CIRUGIA ESPAÑOLA.
Clave: A Volumen: 54.Nº3 Páginas, inicial: 201 final: 203 Fecha: 1993
Lugar de publicación:

Autores (p.o. de firma): J.M. Vallejo, J.L. Vicente Villardón, **M.P. GALINDO**.
Título: Fiabilidad de la codificaicón de estudios meta-aníliticos desde una perspectiva multivariante.

Ref. ☐ revista: REVISTA ESPAÑOLA DE FARMACOECONOMIA ☐ Libro
Clave: A Volumen: V Páginas, inicial: 30 final: 38 Fecha: 1996

ST - 305

Editorial

1019

Metaanálisis: el uso de los métodos estadísticos en revisiones de estudios de investigación relacionados

M.P. Galindo Villardón
Departamento de Estadística y Matemática Aplicadas. Universidad de Salamanca.

Introducción

Uno de los principios fundamentales de la investigación científica es que debe tener la oportunidad de ser replicable. Los resultados de una investigación aislada no se pueden elevar a la categoría de hechos científicos hasta que no hayan sido contrastados por otros investigadores.

El considerable trabajo invertido en la realización de una investigación aporta tan sólo una pequeña pieza de un enorme rompecabezas; es preciso retomar investigaciones anteriores introduciendo variaciones en el procedimiento para poner a prueba la consistencia o la robustez de un determinado hecho empírico.

Este carácter acumulativo del conocimiento científico requiere que entre la investigación pasada y la futura se incluya una etapa intermedia imprescindible para su progreso: *la revisión de la literatura y la resolución de las contradicciones empíricas.* Dado su importante papel, las revisiones de investigación han suscitado un creciente interés en un gran número de ciencias y muy especialmente en medicina.

Su nombre más conocido es *metaanálisis*[1] aunque otros autores lo denominan revisión cuantitativa, integración de investigaciones o análisis cuantitativo de dominios de investigación. Dickersin et al han publicado recientemente un artículo[2] sobre la necesidad de estandarizar la terminología para facilitar la búsqueda de bibliografía. La National Library of Medicine, desde 1989, utiliza el término metaanálisis.

Lo esencial de esta transformación metodológica proviene de la introducción de los métodos estadísticos en el proceso de revisión

Aunque los primeros intentos de desarrollar métodos cuantitativos para integrar resultados de investigación se remontan a los años treinta con Fisher y Pearson, esta metodología no había despertado demasiado interés. Sin duda, son los trabajos de Glass et al los que han impulsado la práctica y, con ella, el desarrollo de las técnicas cuantitativas de integración, cada vez más sofisticadas. Fruto de este interés es la publicación de varios textos de metaanálisis en los últimos años, entre ellos, cabe destacar[3-7] el interesante "Etapas de un metaanálisis".

Rosenthal en su trabajo[4], fijándose en el paralelismo que debe existir entre una investigación primaria y un metaanálisis, diferencia seis etapas básicas por las que debe discurrir cualquier síntesis de investigación:

1. Formulación del problema.
2. Búsqueda de la literatura.
3. Codificación de los estudios.
4. Medida de los resultados.
5. Análisis e interpretación de resultados.
6. Publicación del estudio.

Cada una de estas etapas obliga al revisor a adoptar decisiones que afectan directamente a los resultados de la revisión y, en consecuencia, son propensas a violaciones de la validez de las conclusiones.

Formulación del problema

Análogamente a como se lleva a cabo en las investigaciones primarias cuando se realice una revisión, sea o no cuantitativa, el primer paso consiste en especificar con la mayor rigurosidad posible el tema concreto sobre el que se va a recoger la información.

En todo estudio metaanalítico deben estar claramente delimitados los objetivos; deben ofrecerse definiciones conceptuales de las variables a fin de que queden claramente especificados los estudios primarios que son admisibles.

Las variables deben ser definidas operativamente para que sea posible relacionar un concepto abstracto con un hecho observable, conscientes de que un concepto definido en términos muy restringidos puede afectar a su generalidad y a las inferencias y un concepto definido demasiado ampliamente puede no tener una representación concreta de la población de estudios, proporcionando una generalización exagerada de los resultados.

Búsqueda de la literatura

El paso siguiente a la definición del problema objeto de investigación consiste en la identificación de todos los estudios que han tratado dicho tópico.

La *población objetivo* de un metaanálisis está compuesta por todos los individuos que el revisor espera poder representar en el estudio. Pero en la práctica resulta materialmente imposible encontrar todos los elementos de una población objetivo. La *población accesible* incluirá a aquellos individuos que el revisor es capaz de localizar.

Es requisito fundamental en un metaanálisis definir la población objetivo y la población accesible, así como determinar en qué medida pueden diferir, ya que si ambas no coinciden, cosa que ocurre en la práctica totalidad de los metaanálisis; esto puede suponer una importante amenaza para la validez de los resultados.

Dos fuentes de información deben ser utilizadas para asegurar la adecuación de la población accesible a la población objetivo:

Fuentes informales: consisten en conectar con expertos en el tema de interés de cara a obtener el acceso a material nuevo o no publicado por parte de investigadores que trabajen en el mismo campo. También pertenecen a este apartado los trabajos presentados en congresos profesionales.

Fuentes primarias: recogen las consultas a bibliografías personales y a revistas de investigación primaria.

Cooper[8] distingue los siguientes tipos fundamentales de técnicas para la localización de estudios:

El procedimiento ascendente, que consiste en localizar investigaciones previas que figuran en listas de referencias de estudios ya disponibles.

El procedimiento descendente mediante el que se localizan trabajos en los cuales se citan los documentos ya localizados, por ejemplo, consultando índices de citas.

Búsqueda computadorizada, que puede cubrir los dos anteriores procedimientos y que es hoy día la más rápida y eficaz.

Las distintas fuentes no son excluyentes entre sí.

Otra de las mayores críticas que pueden hacerse al metaanálisis reside en la constatación del *sesgo de selección editorial* a favor de estudios con resultados significativos.

El problema del sesgo de publicación ha sido abordado matemáticamente por Rosenthal y por otros autores. La solución que proponen no es definitiva pero es la que actualmente se utiliza para estimar la tolerancia (*fail-safe number*) de un estudio metaanalítico a resultados nulos; es decir, para determinar el número de resultados nulos que deben existir en los archivos editoriales para alterar los resultados significativos de un metaanálisis.

En el informe de cualquier metaanálisis debe aparecer información exhaustiva en relación a las fuentes de información analizadas, los años cubiertos en la búsqueda y las palabras clave utilizadas en la selección. Sólo de esta forma el lector podrá conocer la fiabilidad de la búsqueda y, por tanto, la validez de los resultados.

Codificación de los estudios

El revisor debe codificar las características de los estudios que supuestamente afecten a los resultados del estudio. Glass distingue entre *características metodológicas* y *carcterísticas sustantivas;* las primeras se refieren a los aspectos generales de la investigación, tales como la fecha y la fuente de publicación, el tipo de diseño de investigación, el tamaño muestral, las características de los sujetos, o la calidad de la investigación. Las características sustantivas son específicas del tópico en estudio.

Este proceso de codificación suele formalizarse elaborando un cuestionario que incluya todas las características relevantes.

El proceso de codificación de las características de los estudios constituye un problema de medida y, por tanto, está sujeto a las deficiencias de fiabilidad y validez del estudio. La falta de fiabilidad en la codificación puede deberse a falta de calidad de los informes o a errores del codificador. Si las reglas de codificación no son lo suficientemente explícitas, distintos codificadores de un mismo estudio pueden realizar juicios diferentes.

Para proteger la fiabilidad del proceso de codificación debe elaborarse un libro de codificación que incluya lo más exhaustivamente posible las definiciones de las características, las normas y las reglas de codificación.

Es recomendable que el proceso de codificación se lleve a cabo por parte de un equipo de codificadores previamente entrenados para la perfecta interpretación y utilización del libro de codificación. Este equipo evaluará una muestra aleatoria de todos los estudios incluidos en el metaanálisis.

Debe llevarse a cabo un estudio sistemático de la fiabilidad intercodificadores; cuanto mayor sea el número de codificadores mayor será la fiabilidad efectiva ya que los errores aleatorios de cada codificador se contrarrestan con los errores de los otros.

Medida de los resultados

En una investigación primaria la unidad de análisis de datos la constituyen las observaciones empíricas, pero en una revisión cuantitativa la unidad de análisis procede de los estadísticos ya analizados en los estudios primarios.

Para poder integrar los resultados de un conjunto de estudios es preciso que se expresen en una escala de medida común. En general, las variables dependientes empleadas en los estudios difieren entre sí, razón por la cual no son directamente comparables.

Este problema se ha intentado solucionar mediante la aplicación de dos estrategias diferentes: *a)* el uso de niveles de significación, y *b)* el uso de medidas de la magnitud del efecto.

El nivel de significación

Si disponemos del nivel de significación exacto "p" que corresponde a cada contraste de significación realizado de una hipótesis común, el conjunto de niveles "p" unilaterales así obtenido es susceptible de ser analizado estadísticamente ya que convierte los resultados de los estudios a una misma medida.

Esta estrategia presenta, sin embargo, varios inconvenientes, quizás el más claro sea que no proporciona información en relación a la magnitud del efecto experimental.

El tamaño del efecto

Green y Hall[9] afirman que la medida más informativa del resultado de una investigación es el tamaño o magnitud del efecto; fue Glass el principal responsable de la introducción de las medidas del tamaño del efecto (TE) en la integración cuantitativa de resultados de investigación.

Cuando la hipótesis nula es falsa, lo es en algún grado específico, es decir, el tamaño del efecto es un valor concreto, distinto de cero, de la población. Cuanto mayor es este valor, tanto mayor es el grado en que se manifiesta el fenómeno bajo estudio.

El nivel de significación estadística sólo permite determinar si un efecto experimental es distinto de cero; el tamaño del efecto permite conocer en qué medida es distinto de cero.

Una amplia gama de indicadores del tamaño del efecto pueden encontrarse en la literatura, tanto paramétricos como no paramétricos (para más detalle, consultar los textos citados en la "Introducción").

Análisis de los resultados

El carácter innovador del estudio metaanalítico alcanza su máxima dimensión en la etapa del análisis de los resultados. Esta aproximación implica una nueva perspectiva del desarrollo del conocimiento mediante la integración de los estudios individuales.

De manera similar a como ocurre en las investigaciones primarias, el metaanálisis exige al revisor que adopte las mismas normas de rigor científico en el análisis e interpretación de un conjunto de resultados de estudios diferentes.

En la etapa de análisis e interpretación de resultados, una vez cuantificados los estudios mediante niveles de significación o mediante estimadores del TE, los resultados son sintetizados para obtener un índice global y representativo del conjunto de estudios. Pero los datos integrados en un metaanálisis deben ser lo suficientemente homogéneos entre sí como para suponer que la medida global es representativa de todos los estudios.

Los primeros trabajos metaanalíticos de Glass et al proponían el uso único y exclusivo de estadísticos descriptivos para sintetizar los resultados globales del metaanálisis. A este respecto, el enfoque exploratorio, tablas del tipo *stem-and-leaf* y gráficos esquemáticos del tipo *Box-plot* resultan especialmente indicados para describir los resultados, pero es evidente que son necesarias más pruebas estadísticas para los estudios metaanalíticos.

Glass señala como única solución las técnicas descriptivas basándose en que los datos procedentes de un metaanálisis no suelen cumplir los requisitos de las pruebas inferenciales tradicionales. Es común encontrar estudios metaanalíticos que aplican las pruebas estadísticas convencionales (ANOVA, análisis de regresión, etc.) como si se tratara de investigaciones primarias. Sin embargo, son probables serias violaciones de las hipótesis de base los estudios metaanalíticos no suelen ser independientes ya que un mismo estudio puede proporcionar varias estimaciones del TE: el supuesto de homocedasticidad raramente se cumple ya que las variancias de los TE son inversamente proporcionales a sus tamaños muestrales.

Técnicas de acumulación de niveles de probabilidad

Los primeros procedimientos estadísticos desarrollados para sintetizar cuantitativamente los resultados de un conjunto de estudios se deben, probablemente, a Fisher y Pearson en los años treinta, quienes desarrollan varias técnicas de acumulación de niveles de significación.

Entre los más utilizados cabe destacar el método de Fisher, el método de Edgington, el método de Winer, el método de Stouffer, el método de Mosteller y Bush, el método de Tippett y el método Logit.

Técnicas de integración del tamaño del efecto

Las técnicas basadas en la combinación de los TE son más informativas que las basadas en los niveles de significación, ya que permiten formular inferencias acerca de la dirección y la magnitud de los efectos.

Para un mayor detalle en torno a los procedimientos de integración en el análisis de resultados, consultar los textos citados en la "Introducción".

Mullen y Rosenthal[10] han desarrollado unos programas de ordenador para contrastar homogeneidad y combinar resultados, aunque el cálculo es tan sencillo que no es estrictamente necesario.

Bibliografía

1. Glass GV. Primary, secondary and meta-analysis of research. Educ Res 1976; 5:3-8.
2. Dickersin K, Higgins K, Meinert CL. Identification of metaanalysis. The need for standard terminology. Cont Clin Trials 1990; 11:52-66.
3. Glass GV, Mc Gaw B, Smith ML. Meta-analysis in social research. Beverly-Hills, Sage. 1981.
4. Rosenthal R. Meta-analytic procedures for social research. Beverly-Hills, Sage. 1984.
5. Hedges LV, Olkin I. Statistical methods for meta-analysis. Londres. Academic Press. London. 1985.
6. Jenicek M. Meta-analyse en Medicine. Evaluation et shynthese de linformation clinique et epidemiologique. St. Hyacinthe et Paris. Edisen et Maloine. 1987.
7. Eddy DM, Hasselblad V, Shachter R. Meta-analysis by the confidence profile method. The statistical synthesis of evidence. Londres. Academic Press, 1992.
8. Cooper HM. Scientific guidelines for conducting integrative research reviews. Rev Educ Res 1982; 552:291-302.
9. Green BF, Hall JA. Quantitative methods for literature Reviews. Ann Rev Psych 1984; 35:37-53.
10. Mullen B, Roshental R. Basic meta-analysis: procedures and programmes. Hillsdale, Erlbaum, U.S., 1985.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 500–515
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Several biplot methods applied to gene expression data

Mira Park[a], Jae Won Lee[b,*], Jung Bok Lee[c], Seuck Heun Song[b]

[a]*Department of Pre-medicine, Eulji University, 143-5 Yongdu-dong, Chung-gu, Daejon 301-832, Korea*
[b]*Department of Statistics, Korea University, 5-1 Anam-dong, Seongbuk-gu, Seoul 701-112, Korea*
[c]*Institute of Human Genomic Study, College of Medicine, Korea University, Gojan1-dong, Danwon-gu, Ansan, Gyeonggi-do, Korea*

Available online 26 June 2007

Abstract

DNA microarray experiments result in enormous amount of data, which need careful interpretation. Biplot approaches show simultaneous display of genes and samples in low-dimensional graphs and thus can be used to represent the relationships between genes and samples. There are several different types of biplots, and these methods need to be evaluated because each plot provides different result.

In this paper, we review several variants of biplot methods such as principal component analysis biplot, factor analysis biplot, multidimensional scaling biplot and correspondence analysis biplot. We investigate the properties of these methods and compare their performances by analyzing various types of well-known gene expression data. We also suggest the supplementary data method as a tool for (i) classifying the previously unknown sample/gene to existing class, (ii) analyzing mixture data and (iii) presenting illustrative variables, etc. The usefulness of this approach for interpreting microarray data is demonstrated.

Keywords: Gene expression data; Biplot; Supplementary data; Principal component analysis; Factor analysis; Correspondence analysis; Multidimensional scaling

1. Introduction

DNA microarray technology has been advanced to the point that it is now possible to monitor gene expression levels on a genomic scale. Currently, two types of microarrays are in common use: 2-channel cDNA microarrays and high-density oligonucleotide microarrays chips such as Affymetrix chips. Every microarray gene experiments result in enormous amount of gene expression data, which need statistical considerations.

Traditional clustering techniques such as hierarchical clustering, k-means clustering and self-organizing map have been applied to the analysis of gene expression data (cf. Eisen et al., 1998; Tamayo et al., 1999; Golub et al., 1999, etc.). It is useful to find gene/sample clusters with similar gene expression patterns for summarizing and interpreting the microarray data. However, it would be more effective if we represent this information by drawing a low-dimensional graph. Visualization of the gene expression data helps us to find and interpret the relationships between genes/samples and to detect outliers. Principal component analysis, often performed by singular value decomposition, has been explored as a method for visualizing large-scale expression data. Raychaudhuri et al. (2000) used PCA to analyze time

* Corresponding author. Department of Statistics, Korea University, 5-1 Anam-dong, Seongbuk-gu, Seoul 136-701, Korea. Tel.: +82 2 3290 2237; fax: +82 2 924 9895.
E-mail address: jael@korea.ac.kr (J.W. Lee).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.06.019

series yeast sporulation expression data. Similarly, Alter et al. (2000) and Holter et al. (2000) analyzed microarray data using SVD. On the other hand, Fellenberg et al. (2001) used correspondence analysis to visualize the relationship between genes and tissues.

In this paper, we review several variants of biplot methods as the visualization tool for exploring gene expression data. Biplot method was originally suggested by Gabriel (1971) and there have been several variants proposed by subsequent researchers (cf. Gower and Hand, 1996). These approaches can show simultaneous display of observations and variables as well as represent the relationships between observations and those between variables in low-dimensional graphs. Here we use PCA (principal component analysis) biplot, FA (factor analysis) biplot, MDS (multidimensional scaling) biplot and CA (correspondence analysis) biplot. We investigate the properties of the resulting graphs and compare the performances of different methods. Also we consider the supplementary data analysis, which is presented by Lebart et al. (1984), for exploratory analysis of microarray data. Several application methods are proposed with illustration of simulated and real data. These methods are evaluated with four well-known data: leukemia data set of Golub et al. (1999), lymphoma data set of Alizadeh et al. (2000), colon cancer data set of Alon et al. (1999) and 60 cancer cell line of Ross et al. (2000).

2. Principles of biplot methods in microarray data analysis

The gene expression data on p genes for n mRNA samples may be summarized by an $n \times p$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level of jth gene in ith mRNA sample. The expression levels might be either absolute (e.g. oligonucleotide arrays) or relative with respect to the expression levels of a suitably defined common reference sample (e.g. cDNA microarrays). Usually, the data are centered (mean zero) and/or standardized (mean zero, variance one) for each gene across the samples.

2.1. Principal component analysis and factor analysis biplot

The singular value decomposition of X is given by

$$X = UDV',$$

where U and V are $n \times r$ and $p \times r$ matrix, respectively, each with orthonormal columns so that $U'U = V'V = I_r$, D is a $r \times r$ diagonal matrix with elements $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r$ in the diagonals, and r is rank of X. Let us define $D^\alpha = \mathrm{diag}(\lambda_1^\alpha, \ldots, \lambda_r^\alpha)$ and let $G = UD^\alpha$ and $H = VD^{1-\alpha}$ where $0 \leqslant \alpha \leqslant 1$. Thus X can be factorized as

$$X = UDV' = GH'$$

for a $n \times r$ matrix G and a $p \times r$ matrix H. And it can be shown that $X_{(s)} = G_{(s)}H'_{(s)}$ provides the best possible rank $s(\leqslant r)$ approximation to X, where $G_{(s)}$ and $H_{(s)}$ are the first s columns of G and H, respectively. One can obtain s-dimensional row (sample) and column (gene) plot by plotting $G_{(s)}$ and $H_{(s)}$, respectively (Gabriel, 1971).

Different values of α lead to different geometries. If we choose $\alpha = 1$, then $G_{(s)} = (\lambda_1 u_1, \ldots, \lambda_s u_s)$ and $H_{(s)} = (v_1, \ldots, v_s)$. And the Euclidean distance between two sample points in the plot represents the Euclidean distance in the complete set since $XX' \approx G_{(s)}G'_{(s)}$. Here the ith row of $G_{(s)}$, consists of the first s principal components for ith sample. We call it principal component analysis (PCA) biplot. On the other hand, if we choose $\alpha = 0$ and take the first s columns of G and H, we have the coordinates for s-dimensional plot. We call this factor analysis (FA) biplot. In FA biplot, cosine between gene points is proportional to the covariance or correlation between genes because $X'X \approx H_{(s)}H'_{(s)}$. In both plots, by superimposing sample and gene plot, we can recover original data since $X \approx G_{(s)}H'_{(s)}$.

2.2. Correspondence analysis biplot

CA was originally developed for 2-way contingency tables (Greenacre and Hastie, 1987). To analyze using CA, the data should be positive number. Thus it is necessary to shift the data additively to be a positive range after centering and standardization before analysis. Now let $X = (x_{ij})$ be the data matrix after shifting, x_{i+} and x_{+j} denote sum of the ith row and jth column, respectively, and x_{++} be the grand total of X. Define $F = (f_{ij})$ where $f_{ij} = x_{ij}/x_{++}$.

The problem can be represented in the form of singular value decomposition as

$$D_r^{-1/2}(F - rc')D_c^{-1/2} = UD_\lambda V',$$

where $r = (f_{1+}, \ldots, f_{r+})'$, $c = (f_{+1}, \ldots, f_{+c})'$, $f_{i+} = \sum_j f_{ij}$, $f_{+j} = \sum_i f_{ij}$, $D_r = \mathrm{diag}(r)$, $D_c = \mathrm{diag}(c)$. U and V are the column orthogonal matrices so that $U'U = V'V = I$, and D_λ is a diagonal matrix whose elements are the singular values of $D_r^{-1/2}(F - rc')D_c^{-1/2}$. We use the first s columns of the matrix

$$A = D_r^{-1/2}UD_\lambda \quad \text{and} \quad B = D_c^{-1/2}VD_\lambda$$

as s-dimensional sample and gene coordinates, respectively. The squared distance between profiles a_i and $a_{i'}$ is given by

$$d^2(a_i, a_{i'}) = (a_i - a_{i'})'D_c^{-1}(a_i - a_{i'}) = \sum_j \frac{(f_{ij}/f_{i+} - f_{i'j}/f_{i'+})^2}{f_{+j}},$$

where $a_i = (x_{i1}, \ldots, x_{ip})'/x_{i+}$. It is called χ^2-distance. In this case, the distances between the points in the plot do not approximate Euclidean distance but approximate chi-squared distance. And we interpret that the genes and samples with similar position correspond to each other.

2.3. Multidimensional scaling biplot

The object of MDS is positioning of the observations into a map such that the interim proximities matched the original dissimilarities (or similarities). There are two essentially different approaches: metric and non-metric scaling methods, each of which has many variants.

Let $E = (e_{ij})$ be squared distance matrix between the rows of X, and define B as

$$B = -\tfrac{1}{2}HEH',$$

where $H = I_n - n^{-1}J_n$ and $J_n = 1_n 1_n'$. The classical metric MDS can be obtained by spectral decomposition of B:

$$B = VD_\lambda V',$$

where $V'V = I_n$ and D_λ is a $r \times r$ diagonal matrix with elements $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r$ in the diagonals, and r is rank of B (Mardia et al., 1979). We can get the s-dimensional sample plot by plotting the first s columns of $VD_\lambda^{1/2}$. Similarly, we can get s-dimensional gene plot for X after defining distance matrix between the columns of X. If we take E as squared Euclidean distance matrix, then the s-dimensional solution gives the same results as PCA biplot. The metric MDS use the actual magnitudes of the original distances to obtain geometric representation.

On the other hand, it is possible to arrange the n observations in a low-dimensional coordinate system using only the rank order of $n(n-1)/2$ original distances and not their magnitudes. When only this ordinal information is used to obtain a geometric representation, the process is called non-metric MDS.

2.4. Interpretation

The distance between row points approximates Euclidean distance of samples for PCA biplot whereas it approximates Mahalanobis distance between samples for FA biplot. From column plot of FA biplot, we can catch the correlation between genes by the cosine of gene vectors, and the coordinates of column points in PCA biplot represent the coefficients for the principal component. The PCA biplot and FA biplot give similar information, but then PCA biplot would be more appropriate when we are more interested in the relationship between samples, whereas FA biplot would be more useful if we want to focus on the relationship between genes. For CA biplot, distance between points is the approximated chi-squared distance in both row plot and column plot. To apply MDS biplot, we have to produce the dissimilarity matrix. For example, we can define the Euclidean distance as a measure of dissimilarity, and then the distance between the points shows the Euclidean distance.

Table 1
Interpretations in various biplot methods

Method	Plot		
	Sample	Gene	Superimposed
PCA biplot	Distance between points: Euclidean distance	Coordinates: coefficient of linear combination	Projection: original data
FA biplot	Distance between points: Mahalanobis distance	Angle between vectors: correlation between genes	Projection: original data
CA biplot	Distance between points: chi-squared distance	Distance between points: chi-squared distance	Matching position: related genes and samples
MDS biplot	Distance between points: dissimilarity (e.g. Euclidean distance)	Distance between points: dissimilarity dissimilarity (e.g. Euclidean distance)	–

For PCA biplot and FA biplot, the projection of row vector and column vector approximates the original data, and we can recover the original data by superimposing the row (samples) and column (genes) plot. In CA biplot, we can interpret the association of genes and samples. For MDS biplot, gene plot and sample plot are produced separately, and superimposing two plots is not meaningful. Table 1 shows the summary of these properties.

3. Supplementary data analysis

It often happens, in practice, that additional information is available that might be added to the original data. Consider we have n_s additional samples and let $Z_+ = (z_{+ij})$ be the added data matrix with n_s rows and p columns. For PCA biplot and FA biplot, supplementary data z_{+ij} should be transformed into

$$x_{+ij} = (z_{+ij} - \bar{x}_j),$$

if the original data x_{ij} are centered. If the original data are centered and standardized, then transformation

$$x_{+ij} = (z_{+ij} - \bar{x}_j)/s_j$$

should be made. Here $\bar{x}_j$ and s_j are the mean and standard deviation of jth variable, respectively. In PCA biplot, from $XV = (UDV')V = UD$, s-dimensional coordinates of additional n_s observations can be obtained by $X_+V_{(s)}$, where $X_+ = (x_{+ij})$ is the transformed data matrix. Similarly, in FA biplot, we plot the first s columns of $X^{+'}U$ for p_s additional variables, where X^+ is the transformed added data matrix with n rows and p_s columns (Lebart et al., 1984).

On the other hand, the supplementary data coordinates for CA are given by

$$a = r^{-1}x_+BD_\lambda^{-1} \quad \text{and} \quad b = c^{-1}x^+AD_\lambda^{-1},$$

where x_+ and x^+ is a new row and column, respectively. It comes from the relationship

$$A = D_r^{-1}XBD_\lambda^{-1} \quad \text{and} \quad B = D_c^{-1}XAD_\lambda^{-1}$$

(Gower and Hand, 1996).

This supplementary data method can be applied to microarray experiments in various situations such as (i) classifying the previously unknown sample/gene to existing class, (ii) analyzing the mixture data, (iii) presenting illustrative variables, (iv) visualization of repeated data, and (v) positioning of outliers, etc. This method can be used to classify unknown genes or samples to the known category with similar expression patterns. Sometimes we have the data obtained from different experimental circumstances, and thus they are not homogeneous, and also we might have some additional variables but those are of a somewhat different nature. For example, we may wish to add prognostic factors such as sex or clinical status to gene expression data. Since the data being analyzed should be homogeneous, it is proper to use supplementary data analysis instead of conducting biplot analysis for whole data. It is also applied for repeated or longitudinal experiments using same samples and genes, and makes interpretation easier and clearer. On the other

hand, if we have several outliers that can distort the whole structure, we can construct the biplot without these samples and then position the outliers as if they are supplementary data. We will examine the usefulness of this method with both simulated and real microarray data.

4. Data analysis

The biplot methods described above are applied to four well-known data sets: the leukemia data set, the lymphoma data set, the colon cancer data set, and the 60 cancer cell line data set. The gene expression data on p genes for n mRNA samples may be summarized by an $n \times p$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level of jth gene in ith mRNA sample. The data were centered for each gene across the samples, but did not need to be standardized to have variance 1 because the measurement scales of the variables in each data are the same. For imputing the missing data, we use k-nearest neighbor algorithm with $k = 5$, in which the neighbors are the genes and the distance between neighbors is based on their correlation (cf. Troyanskaya et al., 2001). To analyze the data by CA, the data should be positive number, and thus we shift the data additively to be a positive range after centering. Since metric MDS biplot with Euclidean distance gives the same results as PCA biplot, we draw non-metric MDS plot rather than metric MDS plot.

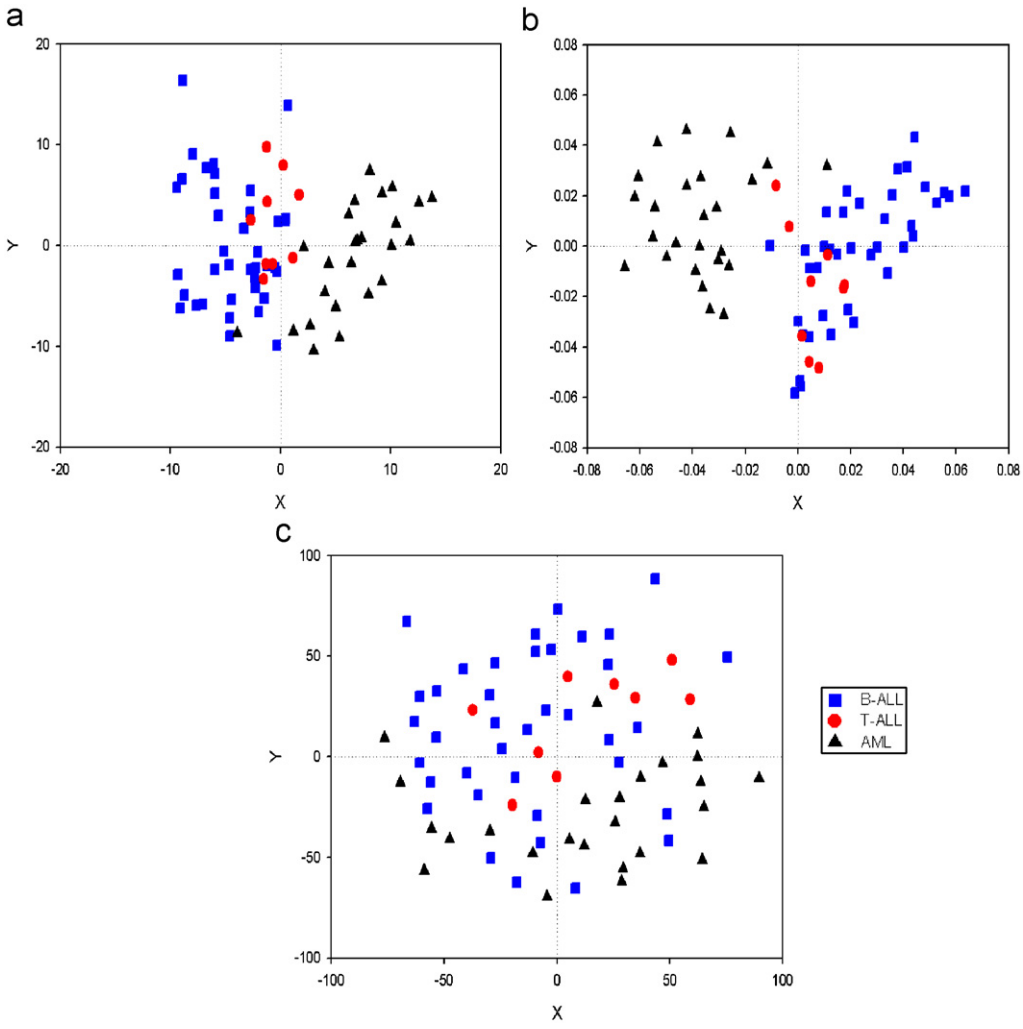


Fig. 1. Leukemia data—sample plot: (a) PCA biplot; (b) CA biplot; (c) MDS biplot.

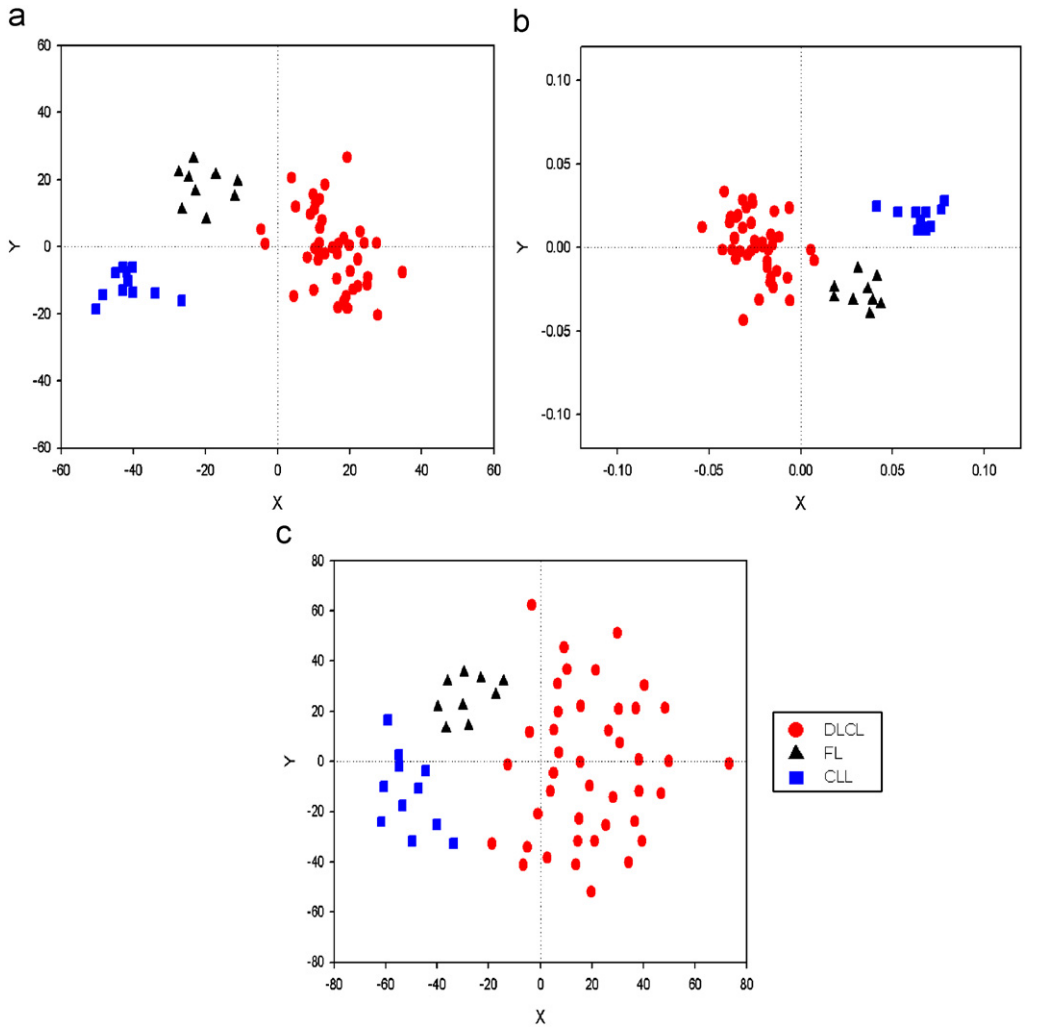


Fig. 2. Lymphoma data—sample plot: (a) PCA biplot; (b) CA biplot; (c) MDS biplot.

4.1. Data sets

Leukemia data: Leukemia data set composed of 3571 gene expressions in three classes of leukemia: 38 cases of B-cell acute lymphoblastic leukemia (ALL), 9 cases of T-cell ALL and 25 cases of acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays (cf. Golub et al., 1999). The data were obtained after preprocessing described in Dudoit et al. (2002). The data can be obtained from http://www.genome.wi.mit.edu/MPR.

Lymphoma data: This data set comes from a study of gene expression of three prevalent adult lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLCL). Among 96 samples, we took 62 samples with 4026 genes in three classes (11 cases of B-CLL, 9 cases of FL, and 42 cases of DLCL). The data matrix consists of the base 2 logarithm of the Cy5/Cy3 fluorescence ratio for gene j in mRNA sample i (cf. Alizadeh et al., 2000). The data can be obtained from http://genome-www.stanford.edu/lymphoma.

Colon cancer data: This data set comes from a gene expression study of 40 tumor and 22 normal colon tissue samples, which were analyzed with an Affymetrix oligonucleotide arrays complementary to more than 6500 human genes. Following Alon et al. (1999), we chose to work with only 2000 genes of highest minimal intensity over the samples. The data can be downloaded from http://www.weizmann.ac.il/mcb/UriAlon.

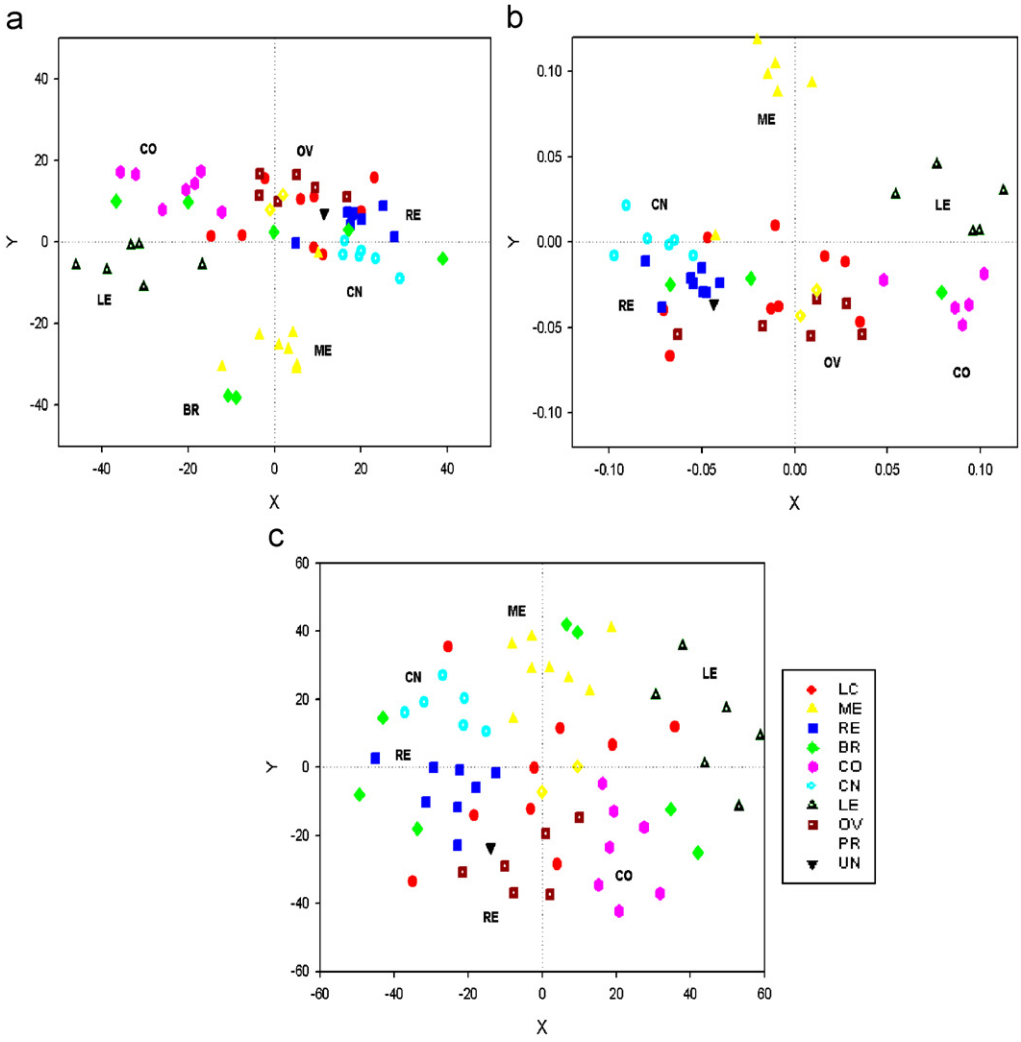


Fig. 3. NCI60 data—sample plot: (a) PCA biplot; (b) CA biplot; (c) MDS biplot.

NCI 60 data: This data set was produced by The National Cancer Institute's anti-cancer drug screen project. The cell lines were derived from various tumor tissues: 7 breast, 5 central nervous system (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 non-small cell lung carcinoma (NSCLC), 6 ovarian, 2 prostate, 9 renal and 1 unknown. The full data set consist of 60 samples and 9703 genes, and we use 1375 genes (Ross et al., 2000). It was studied using cDNA microarrays and the data matrix consists of base 2 logarithm of the Cy5/Cy3 fluorescence ratio. The data can be obtained from http://genome-www.stanford.edu/nci60.

4.2. Analysis using biplots

Sample plot: In sample plot, two nearly located points show that they have similar gene profiles. It shows the relationship between samples and also separates different types of samples. For leukemia data, the first axis of sample plot of PCA biplot and CA biplot separate ALL and AML very well in 2-dimensional plot (Fig. 1(a) and (b)). And T-cell ALL samples are grouped near the origin than B-cell ALL samples. In MDS plot, however, the samples are more widely spread and thus we cannot distinguish each group (Fig. 1(c)). For sample plot of Lymphoma data, three classes are clearly separated in all the biplots (Fig. 2), though the points in MDS plot tend to be dispersed than the others. The first axis separates DLCL cells and the others, and the second axis separates FL cells and CLL cells.

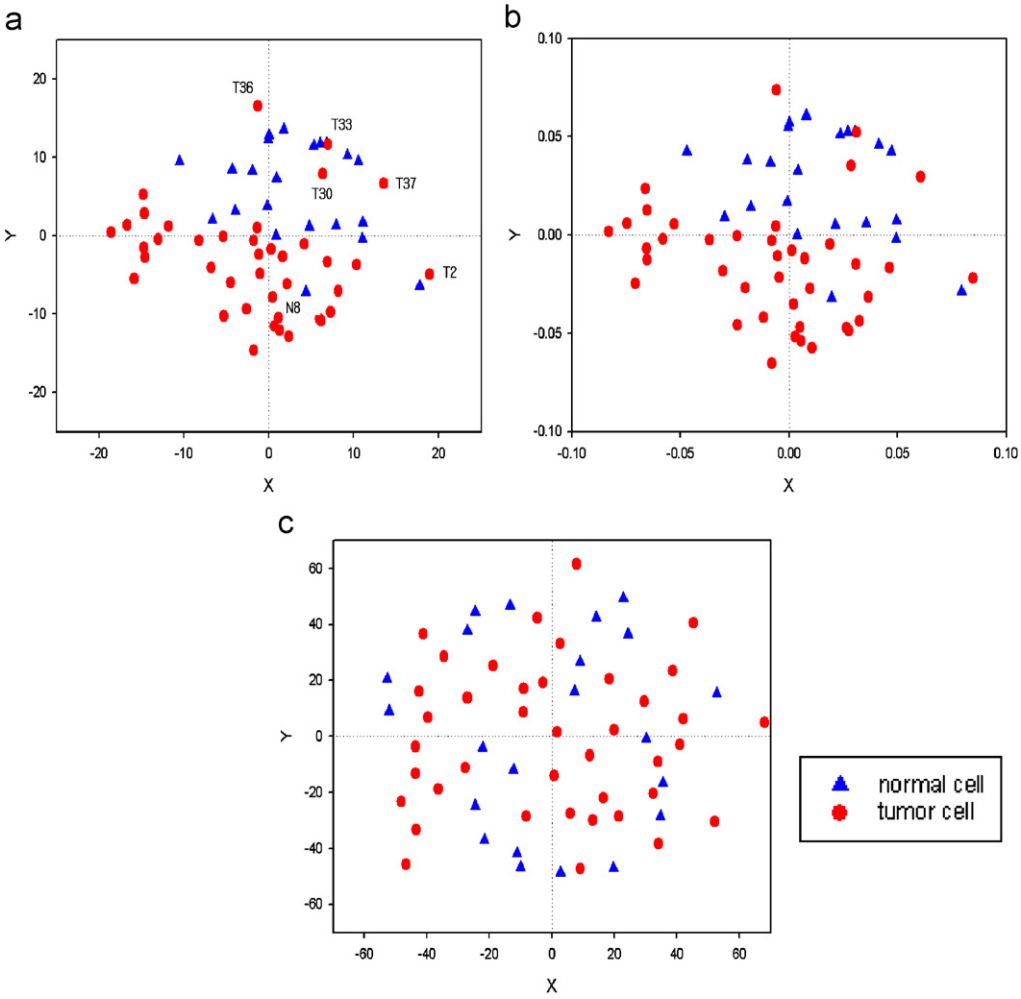


Fig. 4. Colon data—sample plot: (a) PCA biplot; (b) CA biplot; (c) MDS biplot.

For NCI60 data, colon (CO), leukemia (LE), CNS (CN) and melanoma (ME) cells are clustered, respectively, but breast cancer cells (BR) are scattered in PCA biplot (Fig. 3). This plot gives similar results with the dendrogram showing average-linkage hierarchical clustering from Ross et al. (2000). The CA biplot gives the similar results, but the MDS plots cannot separate the clusters.

In PCA sample plot and CA biplot of colon data, the second axis separates tumor cells and normal cells well except for several cells such as T30, T33, T36, T37 and N8. But the MDS plots cannot separate samples and clusters (Fig. 4).

Sample–Gene plot: For both PCA biplot and FA biplot, we can see the association between genes and samples by superimposing row and column plot. The x_{ij} value is big if the jth gene vector lies in the similar direction to the ith sample vector, and the x_{ij} value is close to zero if the jth gene vector is nearly orthogonal. For example, the genes numbered 1–4 in Fig. 5 have large values for DLCL cells such as DLCL0002 and DLCL0026. Alizadeh et al. (2000) defined these genes as "lymph node" signature genes. The genes numbered 5–6 lie in the similar direction to FL cells such as FL10 and FL11, but are nearly orthogonal to CLL cells such as CLL71 and CLL68. It means that these genes are highly expressed in FL cells but have small expression values in CLL cells, and thus they play a role of separating FL and CLL cells. On the other hand, the genes numbered 7–8 are positively related with FL cells and the genes numbered 9–10 are closely related with CLL samples.

We can also interpret the relationships between the genes and the samples from CA biplots. The genes and the samples with similar position are closely related with each other. For example, Fig. 6 shows the plot for CA of leukemia

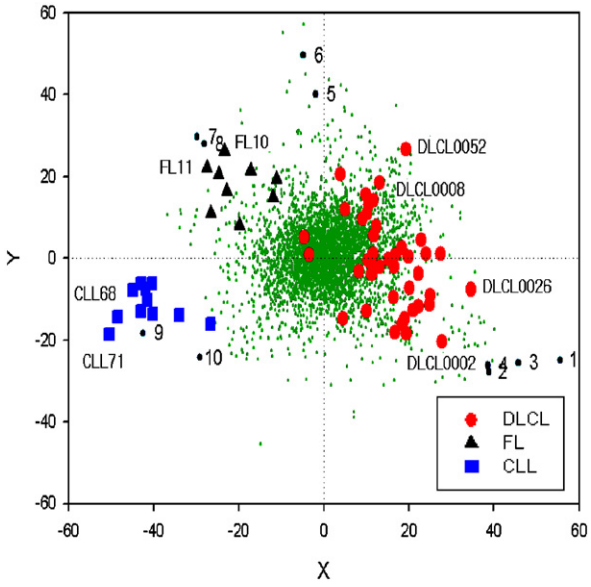


Fig. 5. PCA biplot of lymphoma data.

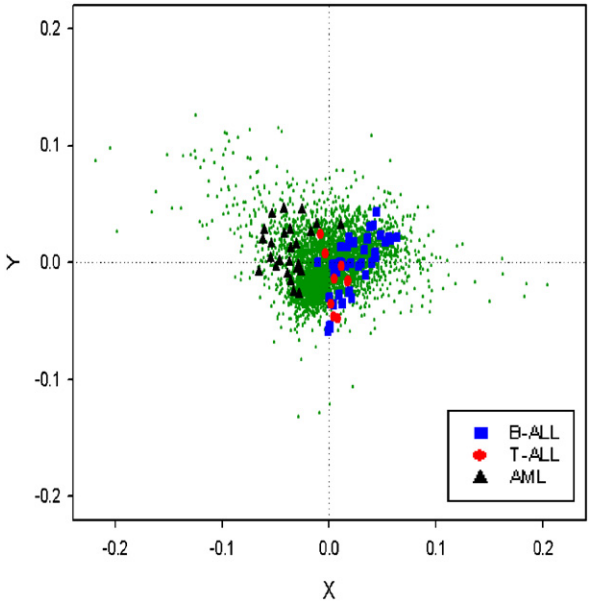


Fig. 6. CA biplot of leukemia data.

data, and the genes plotted on the right side are related with ALL cells and the genes plotted on the left side are related with AML cells.

Gene plot: For the gene plot of PCA biplot, column coordinates are the coefficients of the variables for the principle components. But for FA biplot, cosines of the angles between column vectors represent covariance or correlations between genes approximately, and thus we can interpret that two genes that lie in the similar direction have high-positive correlation. For example, in Fig. 7(a), n1016 has high-positive correlation with n1360 ($r = 0.79$), n1354 ($r = 0.81$), n239 ($r = 0.79$), and they are in the similar direction from the origin. By superimposing sample plot in

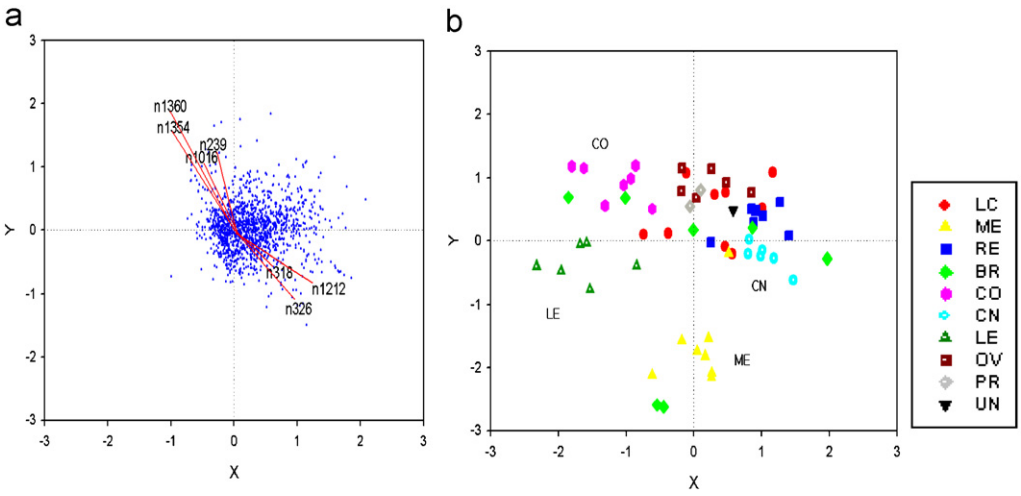


Fig. 7. FA biplot of NCI60 data-gene plot: (a) gene plot; (b) sample plot.

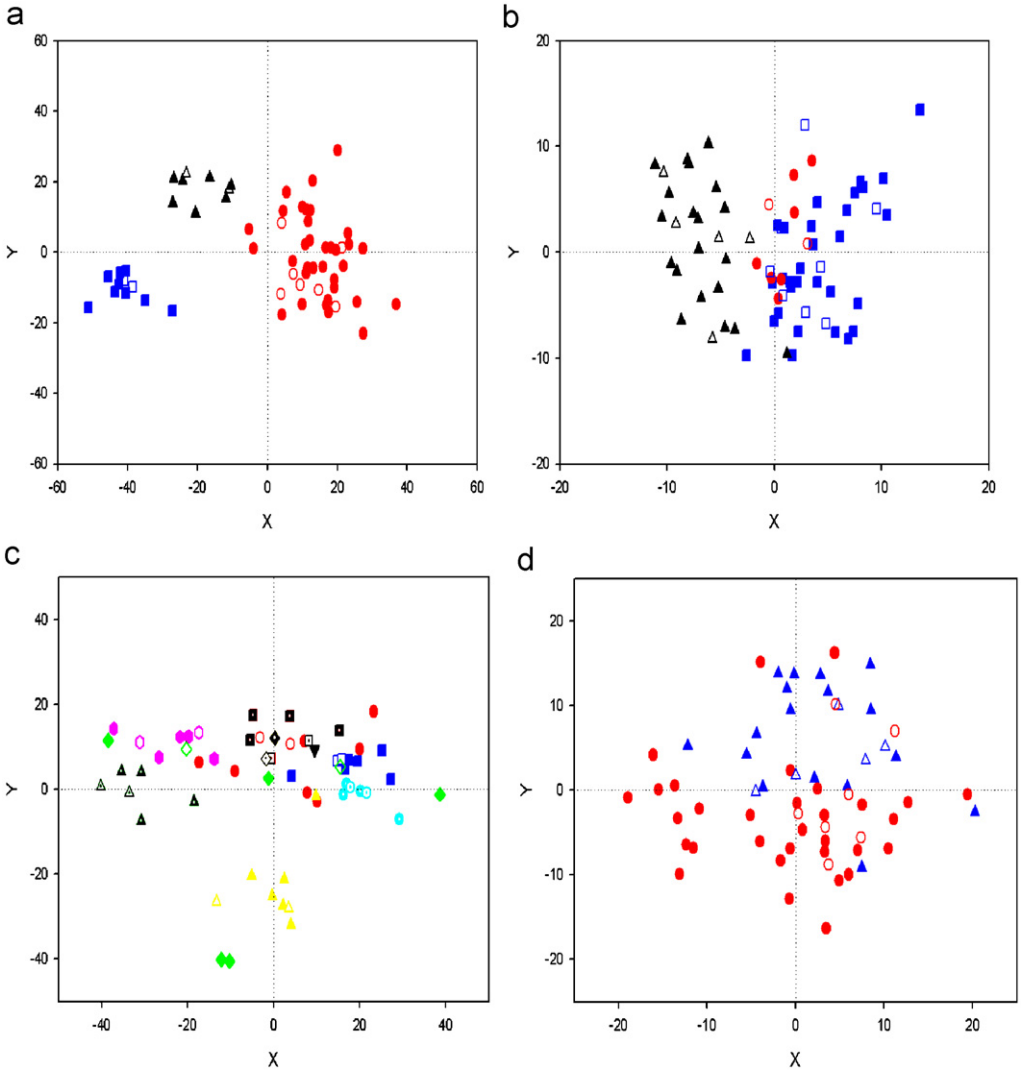


Fig. 8. Simulated supplementary data analysis in PCA biplot: (a) leukemia data; (b) lymphoma data; (c) NCI60 data; (d) colon data.

Fig. 7(b), they seem to be closely related with colon samples (CO). On the other hand, we can see that n1016 has high-negative correlation with n318 ($r = -0.61$), n1212 ($r = -0.60$), n326 ($r = -0.57$), and they lie in the opposite direction in the graph.

4.3. Supplementary data analysis

The supplementary data method can be applied in various situations. We now show three kinds of cases among them as examples.

Case 1 (classification of unknown samples): Supplementary data method can be used to classify unknown genes or samples to the known category with similar expression patterns by treating the unknown genes or samples as additional data. To examine the prediction effectiveness of the supplementary data plot method, we randomly chose 80% of the original data and drew a PCA sample plot. And then we added 20% of the data as a new data. In Fig. 8, the original data are represented by solid diagram whereas the empty one represents the added cells. The shape (color) represents their genuine group. We can see that the additional data are well classified for all the data sets, and plotting the new individuals by supplementary data method is helpful to classify the new samples to the known category. It can also be applied to discriminate new genes to known clusters.

Case 2 (analysis of mixed samples): If we have to handle the mixed data produced from different circumstances, we can apply supplementary data method to overcome the inhomogeneous property. For example, the colon data are the mixture of matched and unmatched samples. Among 62 samples, 44 (22 pair) are matched samples of tumor cell and normal cell, and 18 are unmatched tumor cell. In this case, instead of analyzing all the data at the same time, we can use supplementary data analysis. Fig. 9(a) shows sample plot of PCA biplot using the matched samples only. Here, the blue triangle represents the normal cell, and the solid red circle represents the tumor cell. We drew lines between the several matched samples, and we can find that matched samples tend to be positioned closely to each other. Note that we could not catch it from Fig. 4, which is the plot of all the data. And then we added the unmatched samples as the supplementary data. The unmatched samples are represented by empty black circle (Fig. 9(b)).

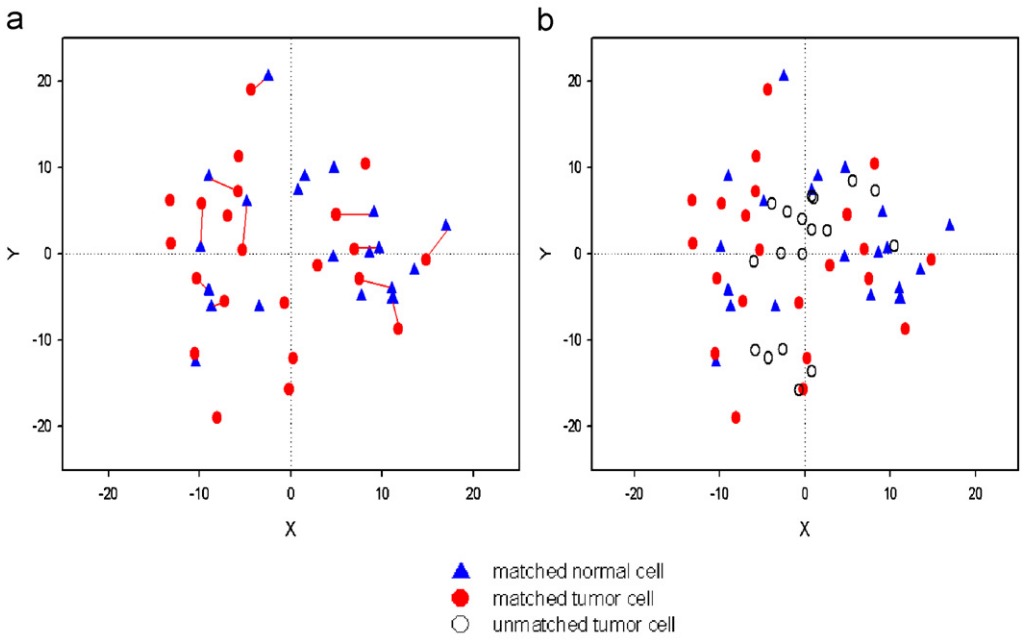


Fig. 9. Supplementary data analysis of matched and unmatched samples from colon data: (a) PCA biplot for matched samples only; (b) supplementary plot adding unmatched samples.

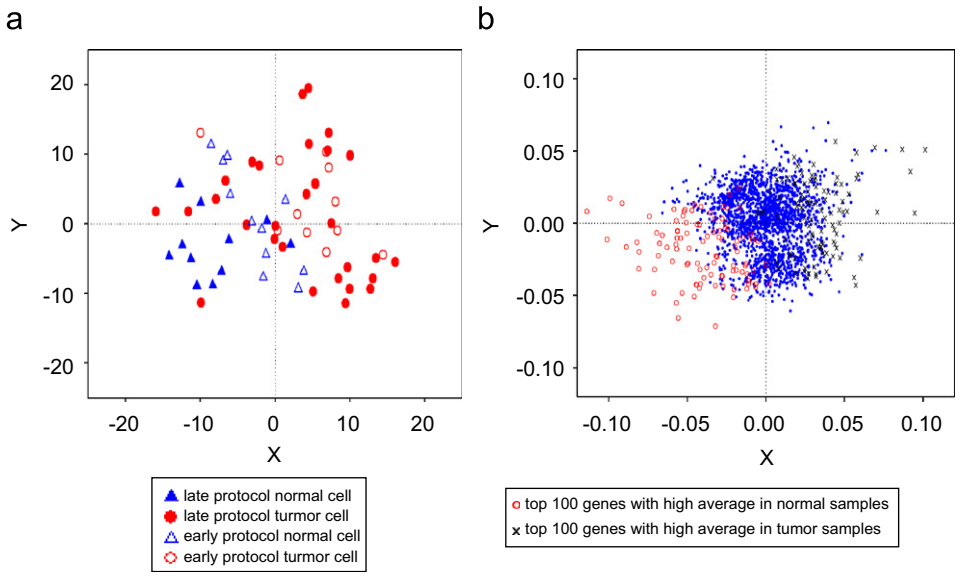


Fig. 10. Supplementary data analysis of early and late protocol samples from colon data: (a) sample plot; (b) gene plot.

On the other hand, there was a change in the protocol for colon data, after the first 11 samples of matched pair data are obtained. And thus we obtained 22 samples by the early protocol and 40 samples by the late protocol. We drew the PCA biplot using the samples adopted by the late protocol first, and then projected the samples adopted by the early protocol onto the map (Fig. 10). The solid one represents the late protocol samples and the empty one represents the early protocol samples. Also the red circle represents the tumor cell and the blue triangle represents the normal cells. In this case, the normal cells tend to lie in the left side of the first axis (Fig. 10(a)), and the genes with relatively higher value in normal samples may lie in the left side of the first axis in gene plot. To check this, we marked top 100 genes with high average in normal samples and tumor samples, respectively (Fig. 10(b)). As expected, we can find the genes with high average in tumor samples (marked as "x") lie in the right side whereas the genes with high average in normal samples (marked as "o") lie in the left side. On the other hand, the tumor cells are spread out and thus they seem to have larger variance compared to the normal cells.

The normal samples proceeded with the early protocol (empty blue triangle) are separated with the late protocol normal samples (solid blue triangle). The early protocol tumor samples also tend to be closer to the origin than the late protocol tumor samples that are located in the left side of the graph, though it is not clearly distinct. Because the early protocol samples are grouped together near the origin, we may interpret that the variations between the tumor cell and the normal cell in the early protocol are smaller than those in the late protocol.

In this example, the experiments can be conducted at two stages. If we have repeated or longitudinal experiment, we can catch the changes in the expression patterns according to the time by constructing the biplot with first data and then by projecting the sequential data sets repeatedly on the first plot.

Case 3 (presentation of illustrative variables): The illustrative variables can be represented by the supplementary data analysis of the gene plot. For example, to represent the early/late protocol and normal/tumor cell, produce $n \times 2$ added data matrix with two new variable x_1^+ and x_2^+. Here, $x_{i1}^+ = 1$ for the early protocol and $x_{i1}^+ = -1$ for the late protocol. Similarly, $x_{i2}^+ = 1$ for the normal cell and $x_{i2}^+ = -1$ for the tumor cell. Then we can apply the supplementary data analysis of FA biplot. Fig. 11(a) shows the plot of two added variable: x_1^+ (normal) and x_2^+ (early). Thus we can interpret that the genes which lie in the similar direction to the black line (NORMAL) tend to be up-regulated in the normal cell, whereas the genes which lie in the opposite direction tend to be up-regulated in the tumor cell. Also the genes which lie in the red line direction (EARLY) tend to be up-regulated in the early protocol cells, whereas the genes which lie in the opposite side tend to be up-regulated in the late protocol cells. Fig. 11(b) shows a snapshot of selecting genes using SAS/JMP. We can identify each gene in the graph and also obtain the list of interesting genes. For example, if we select a set of genes by using lasso or brush tool, then the data for the genes are checked automatically. Also,

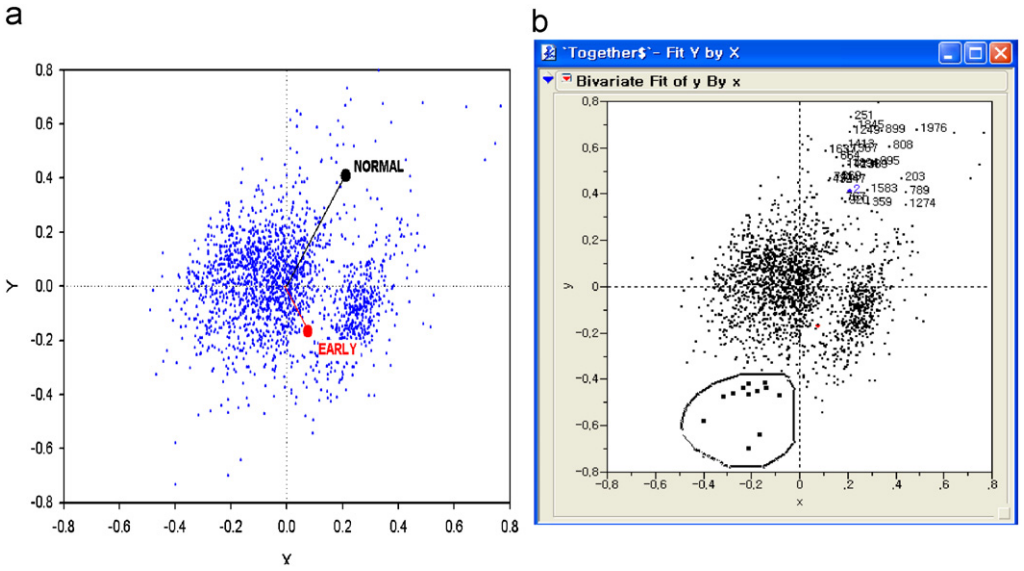


Fig. 11. Supplementary data analysis with illustrative variables in FA biplot of colon data: (a) gene plot with illustrative variables; (b) gene identification using SAS/JMP.

we can label the selected genes. Fig. 11(b) shows the process to label the genes (upper side in the graph) and draw the areas (lower side in the graph).

5. Conclusion and discussion

Application of the several biplots to the publicly available data reveals that the biplot methods can be successful for classifying the samples and exploring the relationship between the genes as well as for overall summarization of the microarray data. While the partitioning clustering methods such as k-means clustering and SOM divided all the data into one of the pre-specified number of subsets, biplot methods enable us to explore the overall aspects of the data by visually inspecting the graph. It is also useful for displaying the genes and the samples simultaneously, and thus the relationship between the genes and the samples can be easily shown in the plot.

Major findings can be summarized as follows. Both PCA and FA biplot perform well for classifying individual, and biplot shows the similar results. The non-metric MDS biplot with Euclidian distance tends to be spread out. From the viewpoint of discriminating the clusters, non-metric MDS biplot performs poorly compared to the other methods for separating clusters, even if sometimes this property is something useful to see the intra-cluster in detail. However, another dissimilarity measure may give the different plots.

We have also shown that the supplementary data analysis might be a useful tool for microarray data analysis in many situations, and thus we strongly recommend to use it. Plotting new samples or genes using the supplementary data method is quite helpful to classify the unknown individual. Moreover it can be applied to deal with the mixture data or outliers. It is also useful to present the demographic variables or the repeated data.

Biplot combined with some partitioning clustering methods such as k-means clustering and SOM can give more meaningful information. One of shortages of partitioning clustering methods is that they do not provide the overall information on the cluster but give information on which clusters the genes belong to. This deterministic nature of partitioning clustering methods can be modified by biplot. Fig. 12(a) shows PCA gene plot with k-means clustering with $k = 10$ for NCI60 data and Fig. 12(b) shows PCA sample plot for lymphoma data when $k = 3$. This plot shows the clustering results graphically. Here, each color shows each cluster, and we can see the k-means clustering gives different suggestions. The FL and CLL cells are clustered in one group (red circle), and DLCL cells are divided into two different clusters (black triangle and blue square). Since we can compare the results by k-means clustering and biplot, these plots can be more helpful to understand the data.

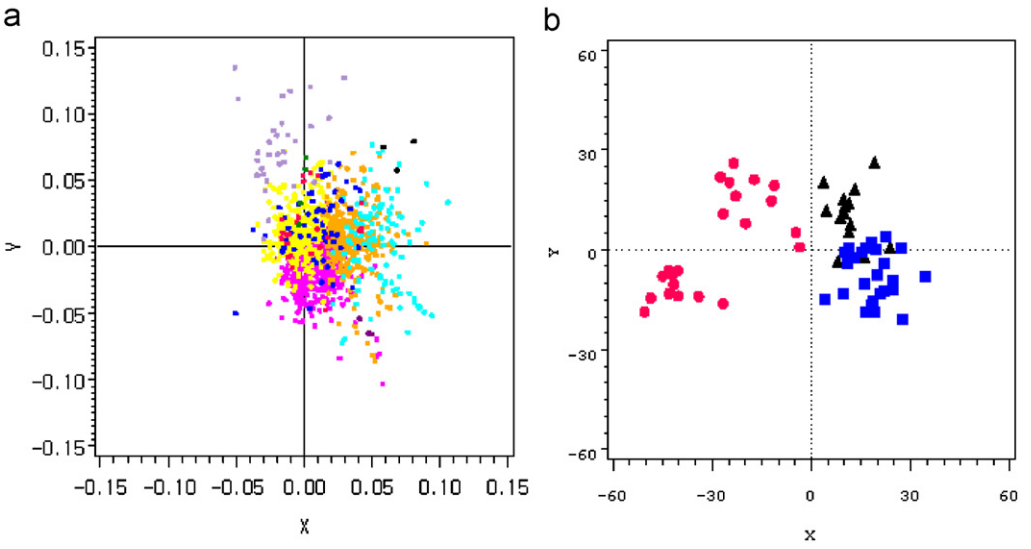


Fig. 12. Biplot with k-means clustering: (a) PCA gene plot of NCI60 data ($k = 10$); (b) PCA sample plot of lymphoma data ($k = 3$).

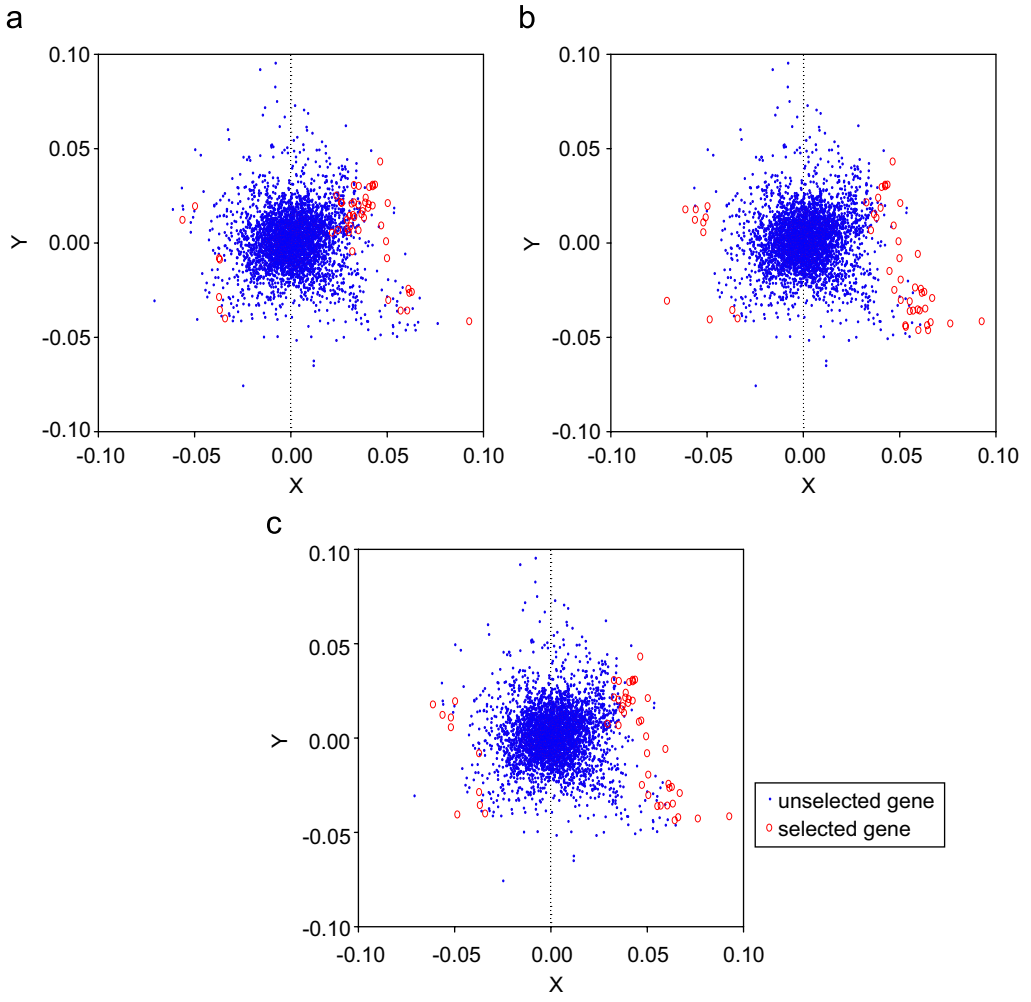


Fig. 13. Gene plot of lymphoma data with several gene selection methods: (a) BSS/WSS criterion; (b) PAM; (c) SAM.

We have used several gene selection methods to select 50 significantly differentially expressed genes. In Fig. 13(a), we have selected significantly differentially expressed genes based on BSS/WSS criterion, which was used in Dudoit et al. (2002). We have also used the soft thresholding methods as in PAM (Tibshirani et al., 2002) and in SAM (Tusher et al., 2001). The selected genes are plotted as red empty circles. As shown in the figure, by applying the biplot method and gene selection method simultaneously, we can easily identify the significantly differentially expressed genes and investigate the relationship between the selected genes simultaneously. It can be a meaningful initial step to explore the gene expression data. Moreover, if we apply biplot methods to the selected gene subset instead of whole genes, we could catch the relationships between the samples and the genes more clearly. In this study, we define the samples as the observations and the genes as the variables. If we apply the methods to the transposed data, we will get the other types of plots and thus the properties of the plot will be changed.

In this study, we have used SAS/IML procedure to obtain the coordinates of the points. For automatic system for pointing out interesting genes/samples, a programming using SAS/JMP or JAVA will be needed.

Acknowledgments

We would like to express our sincere thanks to two referees and the coordinating editor for their helpful and constructive comments. Mira Park and Seuck Heun Song were supported by the Korea Research Foundation Grant founded by Korean Government (R14-2003-002-01000-0). Jae Won Lee was supported by the Korea Research Foundation Grant founded by Korean Government (2005-070-C00020).

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J.J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Bostein, D., Brown, P.O., Staudt, L.M., 2000. Different type of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403, 503–511.

Alon, U., BarKai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci. 96, 6745–6750.

Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Nat. Acad. Sci. 97 (18), 10101–10106.

Dudoit, S., Fridlyand, J., Speed, P., 2002. Comparison of discrimination methods for classification of tumors using gene expression data. J. Amer. Statist. Assoc. 97, 77–87.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Clustering analysis of display of genome-wide expression patterns. Proc. Nat. Acad. Sci. 95, 14863–14868.

Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J., 2001. Correspondence analysis applied to microarray data. Proc. Nat. Acad. Sci. 98 (19), 10781–10786.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453–466.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfiled, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Gower, J.C., Hand, D.J., 1996. Biplots. Chapman & Hall, London.

Greenacre, M., Hastie, T., 1987. The geometric interpretation of correspondence analysis. J. Amer. Statist. Assoc. 82, 437–447.

Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., Fedoroff, N.V., 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc. Nat. Acad. Sci. 97 (15), 8409–8414.

Lebart, L., Morineau, A., Warwick, K., 1984. Multivariate Descriptive Statistical Analysis. Wiley, New York.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

Raychaudhuri, S., Stuart, J.M., Altman, R.B., 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pacific Symposium on Biocomputations, pp. 455–466.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O., 2000. Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. 24 (3), 227–235.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Nat. Acad. Sci. 96, 2907–2912.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Nat. Acad. Sci. 99, 6567–6572.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525.

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarray applied to the ionizing radiation response. Proc. Nat. Acad. Sci. 98, 5116–5121.